Educational leadership effectiveness: a Rasch analysis

Claire Sinnema
Faculty of Education, The University of Auckland, Auckland, New Zealand
Larry Ludlow
Lynch School of Education, Boston College, Boston, Massachusetts, USA, and
Viviane Robinson
Faculty of Education, The University of Auckland, Auckland, New Zealand

Received 16 December 2014
Revised 5 June 2015
4 October 2015
Accepted 12 October 2015

Abstract

Purpose – The purposes of this paper are, first, to establish the psychometric properties of the ELP tool, and, second, to test, using a Rasch item response theory analysis, the hypothesized progression of challenge presented by the items included in the tool.

Design/methodology/approach – Data were collected at two time points through a survey of the educational leadership practices of school principals ($n = 148$) and their teachers ($n = 5{,}425$). The survey comprised seven effectiveness scales relating to school-wide dimensions of leadership, and one scale relating to the effectiveness of individual principals' leadership. The authors undertook validation of the hypothesized structure of the eight ELP scales using the Rasch rating scale model.

Findings – The authors established constructs that underpin leadership practices that are more and less effectively performed and determined the nature of their progression from those that are relatively routine through those that are more rigorous and challenging to enact. Furthermore, a series of analyses suggest strong goodness-of-model fit, unidimensionality, and invariance across time and educator group for the eight ELP scales.

Research limitations/implications – This study focussed on experienced principals – future studies could usefully include school leaders who are new to their role or compare leadership patterns of higher and lower performing schools. A useful future direction would be to investigate the predictive validity of the ELP tool.

Originality/value – This study reveals the ELP is a useful tool both for diagnosing leadership effectiveness and, given that it is essentially stable over time, may prove useful for charting the effectiveness of leadership development interventions.

Keywords Principals, Leadership, Leadership development, Educational administration, Rasch

Paper type Research paper

With confirmation of the central role of school leadership in the performance and improvement of schools (Robinson *et al.*, 2008; Orr and Orphanos, 2011) there is an increased focus on the need for high-quality measurement of leadership effectiveness. It is widely agreed that current leadership evaluation practices are limited by poorly validated tools (Scherbaum *et al.*, 2006; Grissom and Loeb, 2011), and evaluation processes that fail to provide leaders with useful and rigorous feedback (Goldring *et al.*, 2009a).

The conceptual frameworks that have traditionally informed leadership measures used in education have been adult centric – that is, they have focussed on the type and quality of leaders' relationships with other adults. Transformational leadership is one form of leadership with such a focus – it attends to how leaders influence other adults and on the quality of relationships between leaders and followers. Transformational leaders in schools seek relationships with teachers that make them feel valued, that encourage teachers' creativity, and that communicate optimism, high expectations and

Journal of Educational Administration
Vol. 54 No. 3, 2016
pp. 305-339

0957-8234
DOI 10.1108/JEA-12-2014-0140

a shared vision. They strive to inspire and motivate through the quality of their relationships. While many studies show leadership of this type to impact on staff attitudes, the evidence suggests the impact of these attitudes on student outcomes is very small (Robinson *et al.*, 2008).

More recently, attention has shifted to leadership approaches that are more student centered. Such approaches seek to identify and evaluate the leadership practices that have been shown to make a difference for student outcomes (Goldring *et al.*, 2009b; Robinson *et al.*, 2008). For example, leaders' goal-setting or direction-setting practices have emerged as important from both qualitative and quantitative reviews of the published evidence about the relationship between types of leadership practices and student outcomes (Leithwood *et al.*, 2008). Similar links to student outcomes have been found for such leadership practices as promoting and participating in teacher learning and ensuring an orderly and safe environment for both staff and students (Heck, 2000; Timperley *et al.*, 2007; Timperley and Alton-Lee, 2008; Vescio *et al.*, 2008). This research on the links between specific educational leadership practices (ELP) and student outcomes is increasingly informing new leadership evaluation tools, many of which take a multi-source approach whereby principals, teachers and principals' supervisors rate the performance of the principal.

An additional consideration in the design of such tools in the New Zealand context is that school leaders are expected to establish learning environments that are responsive to the educational and cultural needs of the indigenous population (Māori), and are required to report separately to their community and government on the academic and cultural achievement of this group. Furthermore, current policy requires responsiveness to other groups of priority learners. A tool was required, therefore, that assessed leaders' effectiveness in ensuring responsiveness to the diverse learners that are found in most New Zealand schools. The ELP tool described in this study was designed primarily to provide New Zealand principals and their leadership teams with formative feedback about the quality of their leadership of learning and teaching and with some diagnostic insights about the practices in which they were perceived to be more and less effective. The purposes of this study are, first, to establish the psychometric properties of the ELP tool, and, second, to test, using a Rasch item response theory analysis, the hypothesized progression of challenge presented by the items included in the tool.

The ELP tool

The conceptual framework for the ELP tool was provided by the findings from a synthesis of outcomes-linked evidence that explains the relationship between school leadership and student outcomes (Robinson *et al.*, 2009; Robinson, 2007). Our study focusses on the construct validity of the ELP tool, rather than investigating its predictive validity in terms of student outcomes. The prior established link between those constructs and positive impacts on learners is, though, central to the justification for the ELP framework.

We make no claim that results from the ELP are therefore predictive of student outcomes – that is the work of a future study, but a framework focussed on leadership practices previously found to influence student learning (even indirectly) is more justified, in our view, than a framework where the links between the constructs and student learning are unknown. For that reason we overview here three sets of findings from the leadership best evidence synthesis (Robinson *et al.*, 2009; Robinson, 2007) that the ELP framework draws on. The first was a meta-analysis of quantitative studies

that investigated the statistical relationship between measures of leadership practices and measures of student outcomes. The findings from that meta-analysis were derived using a forward-mapping strategy to examine the evidence of the impact of leadership on school conditions that indirectly impact on student outcomes. Those findings indicated the relevance of the first six school-wide leadership dimensions in the ELP framework: goal setting, strategic resourcing, curriculum quality (CQ), quality of teaching, teacher development, and safe and orderly environment. We use the term "school-wide" to signal that these dimensions reflect the work of the whole school leadership team and not just the principal. The seventh school-wide leadership dimension in the ELP framework – families and community – was included since it was indicated in the second set of findings in the fore-mentioned synthesis. Those findings were derived using a backward-mapping strategy that examined how interventions in teacher professional learning, Maori-medium settings, and school-community partnerships impacted on school conditions that indirectly impacted positively on student outcomes. The backward-mapping strategy took impact on students (resulting from the interventions) as the starting point, from which implications for school leadership were derived or inferred. The findings indicated the potential of leadership practices focussed on educationally powerful connections between home and school to contribute positively to student outcomes. A further meta-analysis examining the effect sizes of various approaches schools take to connecting with families revealed both the potential impact of such connections (overall effect of 0.42) and the variability in the impact of the different approaches. The eighth dimension of the ELP framework – principal leadership (PL) – was also drawn from the Robinson *et al.* (2009) leadership synthesis. It draws on findings about the leadership capabilities that make a difference for student outcomes including pedagogical knowledge and skills and dispositions required for complex problem solving, challenging conversations and building relational trust. Findings about those leadership capabilities were established through a predominantly backward-mapping approach – since findings directly linking leadership capabilities to student outcomes are scarce, the synthesis looked to findings about the capabilities linked to the other dimensions for which there was evidence of impact on students. The eighth ELP dimension for PL is distinct from the prior seven dimensions – it requires responses about the capabilities of an individual principal whereas responses to items for the first seven dimensions rate the effectiveness of the school leadership team as a whole.

Items for the eight scales of the ELP were then developed using the framework dimensions outlined above. In addition, a New Zealand Ministry of Educational Leadership Framework was used to ensure the inclusion of items about responsiveness to Māori and other diverse learners within each scale (Ministry of Education, 2008). The ELP was developed by the New Zealand Council for Educational Research under contract to the New Zealand Ministry of Education. The third author was also involved at the stage of item writing. Between eight and 16 items were written for each dimension. The stem used for the school-wide leadership dimensions was the same for both principal and teacher respondents – "How effective is the leadership of your school in ensuring that […]" The stem for the PL dimension was adapted for the two types of respondent. Principals were asked "How effective are *you* in […]" while teachers were asked "How effective is *the principal of your school* in […]."

The items were written to include a range of "difficulty" with deliberate inclusion of aspirational items likely to show change over time. Difficulty, in our context, refers to the extent to which a leadership practice item is relatively easier or harder for

principals and schools to be rated on as "outstandingly effective." The tool was designed for online completion by principals as well as the teachers at their school. The combination of self and other ratings enables principals to receive a report that compares their ratings of their own leadership and of school-wide leadership with their teachers' ratings on each of the scales and for each item within each scale. The ELP tool thus provides 180-degree feedback to principals about the match between their own and their teachers' perceptions of the effectiveness of school-wide and PL.

Three processes were carried out to pre-test the questionnaire. The first involved a combination of a form of behavior coding and cognitive interviews (Presser and Blair, 1994). The purpose of that process was to establish the cognitive validity of the ELP tool. Cognitive validity testing provides evidence of the alignment between respondents' thoughts, beliefs and feelings in response to questionnaire items, with the intended outcomes of the instrument (Karabenick *et al.*, 2007). An educational leadership policy group including leadership practitioners and professional development providers, all with teaching and/or school leadership experience, were asked to respond to each of the questionnaire items. A record of all items for which they needed to ask for clarification or indicated some degree of difficulty in responding was kept. The group then shared retrospective think-alouds for those items to indicate how they had interpreted the question overall and the meanings they associated with key terms in the item. The group, led by the questionnaire design team, then discussed the alignment between those interpretations and the intent of the item. On the basis of their responses multiple items were revised. The second process involved an expert panel of four academics with experience in questionnaire design and administration in the context of educational leadership research. They were asked to provide feedback in relation to both the overall structure of the questionnaire, individual items and the instructions to principals and teachers. Their feedback led to increased consistency of item wording, and ensured that important conceptual omissions and ambiguous wording were addressed. The final process involved the trialing of the questionnaire in 36 volunteer schools. Following completion of the questionnaire trial participants were asked to provide feedback on their experience of the electronic administration of the questionnaire.

In the following section we describe the conceptual framework that guided item writing for these eight dimensions.

School-wide leadership practice dimensions

For each of the seven school-wide dimensions, we describe the conceptual and empirical basis of the dimension and the nature of the items included in each scale. The items within each scale were intended to vary in difficulty so that some would present more challenge than others in terms of leadership practice.

Goal setting. Establishing goals and expectations, and communicating those goals in ways that gain the commitment of those responsible for achieving them is important leadership work (Leithwood *et al.*, 2008; Seijts and Latham, 2012; Hallinger and Heck, 2002; Heinrich, 2012). The 11 goal-setting items in the ELP tool focus on the extent to which schools' strategic goals and targets promote high standards and expectations for all students, on the extent to which they are based in evidence of student needs, on how the goals are communicated, and on how progress toward them is monitored. Our hypothesis was that items about setting goals would be easier to rate as outstandingly effective than those that stipulate rigorous inquiry into evidence for the setting and evaluation of goal achievement.

This hypothesis was based, in part, on the fact that while New Zealand principals are required to set goals for their own and their school's development, the quality of their goal setting and analysis is not high (Sinnema and Robinson, 2012).

Strategic resourcing. In order for goals to be achieved, appropriate resources, including money, time and people, are required to help meet those goals (Miles and Frank, 2008; Grissom, 2011). Strategic resourcing items in the ELP focus on the organization of teaching resources, timetables, routines and the use of expertise. Our hypothesis was that items about resourcing for learners generally would be easier to rate as more highly effective than those about resourcing for groups of learners with particular needs. Once again, this reflects the New Zealand context in which provision for learners with special needs is not governed by special legal requirements as in the USA. In addition, the New Zealand school system has a long history of under-serving its indigenous (Māori) and immigrant and New Zealand-born Pacific communities (May *et al.*, 2012).

Curriculum Quality (CQ). When school leaders take an active and developmental role in planning, coordinating and evaluating the quality of the curriculum and programs that guide teaching, students in those schools are more likely to achieve well (Robinson, 2011; Grissom *et al.*, 2013). The CQ items in the ELP relate to the relevance of curriculum content to various learners, to the level of challenge in programs that students experience and the attention to evidence about learning and goal achievement when planning school-based curricula. Our hypothesis for the items in this scale was that those focussed on the more administrative aspects of dealing with CQ (ensuring plans are in place, for example) would be easier to rate as outstandingly effective than those requiring high-quality curricula for all learners in all learning areas.

Quality of teaching. As well as ensuring that the curriculum is of high quality, leadership should also be directly involved in planning and evaluating the quality of teaching. This requires active oversight and coordination of teaching and learning programs, leadership of discussions about instruction and its impact on students, observations of teaching followed by developmental feedback, and systematic monitoring of student progress (Heck *et al.*, 1990; Heck *et al.*, 1991; Robinson, 2011). Those practices were captured in the quality teaching items in the ELP through, for example, items about improving teaching, identifying teaching difficulties, focussing teacher evaluation on improvement, the use of data and feedback on teaching effectiveness. Our hypothesis here was that items focussed on practices with the potential to be considered as collegial (about helping, supporting and sharing responsibility) might be easier to rate as outstandingly effective than those emphasizing more rigorous progression practices (such as improvement, discussion of problems, feedback and challenge). This hypothesis reflects literature about the capabilities required of principals in order to effectively carry out teacher evaluation (Sinnema and Robinson, 2007).

Teacher development. Much empirical evidence supports the idea that "the most powerful way that school leaders can make a difference to the learning of their students is by promoting and participating in the professional learning and development of their teachers" (Robinson, 2011, p. 104). This includes both formal and informal opportunities for development and requires school leaders to be an accessible and knowledgeable source of instructional advice (Friedkin and Slater, 1994; Grissom *et al.*, 2013). The ELP items relating to leadership that promotes and participates in teacher learning refer to the analysis of achievement data for planning professional learning, to serious discussion as a means of developing teaching quality, and to the role of evidence in evaluating the

effectiveness of improvement efforts. We hypothesized with regard to the items in this dimension that those requiring engagement with evidence (including data about their own students' progress) might be harder to rate as outstandingly effective than those items without an emphasis on evidence. This hypothesis reflects the international and New Zealand evidence about the challenge of using data for improvement purposes (Datnow and Park, 2014; Education Review Office, 2013).

Safe and orderly environment. This dimension includes those management tasks that ensure the smooth functioning of the school and a secure learning environment for (and as perceived by) both staff and students (Wang and Holcombe, 2010; Robinson, 2011; Heck *et al.*, 1991). There is some evidence that principal effectiveness in such tasks has small but statistically significant effects on student achievement (Grissom and Loeb, 2011). In the ELP tool, items about this dimension of leadership ask about school safety, orderliness, and the ability of the leadership to resolve problems and conflict effectively. Our hypothesis here was that items focussed on routine management and monitoring would be easier to rate as outstandingly effective than those involving resolving problems in relation to the school environment.

Connections with family and community. The role of creating educationally powerful connections between school and home and between feeder schools is a vital one (Robinson *et al.*, 2009). To be educationally effective, the connections should have an explicit focus on enhancing student learning, rather than on fundraising or governance (Borman *et al.*, 2003). They should also promote high levels of trust within the school community, since there is strong evidence about the relationship between the level of trust within a school community and improvement in student achievement (Bryk and Schneider, 2002). ELP items for this dimension emphasize the responsiveness of schools to family views about teaching and learning, and to the quality of interaction, partnership and communication between home and school. They are underpinned by an understanding of the relationship between the level of trust within a school community and improvement in student achievement. Our hypothesis for this dimension was that items requiring the provision of information would be easier to rate as outstandingly effective than those requiring more two-way interactions and genuine partnership between home and school or between organizations. This hypothesis reflects evidence about the challenge of ensuring that teachers do not work in isolation from, but rather in partnership with, other influential people in children's lives (Epstein and Sheldon, 2006).

The dimension of Principal Leadership (PL)

The PL dimension in the ELP comprised items focussed on an individual principal, rather than on a school-wide leadership team as in the previous seven dimensions. The focus is on the capabilities, personal qualities and interpersonal skills that are required or implied by the leadership practices described in the prior school-wide leadership dimensions (Robinson, 2011). The items are also based in empirical research on the determinants of teacher and parental trust of principals (Bryk and Schneider, 2002; Goddard *et al.*, 2009). They ask, for example about how effectively the principal resolves conflict, learns alongside teachers, earns the respect of the community, shows personal and professional respect for staff, and is open to learning and admitting mistakes. Our hypothesis for this dimension was that items describing skills required to address interpersonal problems and conflict effectively would be harder to rate as outstandingly effective than those about more general qualities (such as respect).

Sample and administration

The principal participants in this study ($n = 148$) were recruited at two different time points. In total, 90 experienced principals were recruited from throughout the greater Auckland region during their attendance at a professional development seminar. An additional sample of 58 principals was recruited a year later, whose characteristics were similar to those in the original sample in terms of principal gender, school type (primary, intermediate, secondary, composite) and school socioeconomic status (low, medium, high). Of the 90 principals in the original sample who participated at time one (Time 1), 67 percent ($n = 60$) also completed the survey at time two (Time 2). In total, 39 of the 58 principals from the supplementary sample (67 percent) completed the survey on both occasions. There was a 15-month time period between the first and second administrations of the survey for both groups. Teachers at the schools of principals who had already agreed to participate were invited by a survey administrator to also participate – 3,162 took part at Time 1, and 2,267 at Time 2.

For principals the response rate for the original sample was 95 percent at Time 1 and 73 percent at Time 2 and for the supplementary sample the response rate was 89 percent at Time 1 and 65 percent at Time 2. The relatively high rates were likely due to participants being offered the incentive of personalized reports that were highly relevant to the requirements for New Zealand principal evaluation and school review processes, and to the expectations placed on school leaders by their governing bodies for data on their leadership practice. The reports, which were provided at no cost, detailed scores for each of the scales described earlier and each individual item in a way that compared principals' own ratings with the collective response of their teaching staff. Principals were offered a book voucher and a professional learning opportunity to discuss the implications of the research findings for their own leadership.

The response rates for teachers were also at acceptable levels – for the original sample the overall response rate was 61 percent at Time 1 and 64 percent at Time 2, and for the supplementary sample the overall response rate was 66 percent at Time 1 and 46 percent at Time 2. Efforts to ensure adequate teacher response rates included careful guidance about how to introduce the purpose and uses of the survey and information about why a high response rate was needed to increase the reliability of information about leadership in the school. Schools were also encouraged to allocate a specific non-teaching or meeting time to enable teachers to complete the survey. Duplicate responses were not possible. For each survey administration, teachers were required to enter their e-mail address and a unique code and were not technically able to re-enter and create another set of survey responses. A manual monitoring process ensured that no unexpected e-mail addresses were used to complete any survey response.

Psychometric properties of the ELP scales

We undertook validation of the eight ELP scales using a Rasch measurement perspective (Ludlow *et al.*, 2014; Rasch, 1960): the items should be unidimensional, they should vary from easier to harder in their difficulty, the spread of item difficulty should be uniform, their easy-to-difficult spread should follow a hierarchical progression, the items should be of equal discrimination, the items should be independent in the sense that an answer to one is not dependent upon the answer to another and item revisions and rejections should be conducted so that the items fit the model. For present purposes, these principles mean that we expected that each of the eight separate scales would consist of a substantively meaningful, unidimensional progression from relatively commonplace and routine to

more complex and demanding practices and principals and teachers would demonstrate agreement on how they perceive these ordered progressions of practices (For similar applications see Sinnema and Ludlow, 2013).

The Rasch model

The Rasch rating scale model (Rasch, 1960; Andrich, 1978; Wright and Masters, 1982) was employed for the analysis of the eight scales at both Time 1 and Time 2. The rating scale model is appropriate when the rating categories (ineffective (1), minimally effective (2), satisfactorily effective (3), highly effective (4) and outstandingly effective (5)) are intended to have the same relative meaning for all items. That is, the understanding of what differentiates an "outstandingly effective" rating from the slightly lower "highly effective" rating is assumed to hold regardless of the specific practice assessed. If the rating options were not intended to function in this manner (e.g. for some practices it is harder to achieve a "highly effective" rating than it is to achieve an "outstandingly effective" rating), then the Rasch partial credit model (Wright and Masters, 1982) would have been employed.

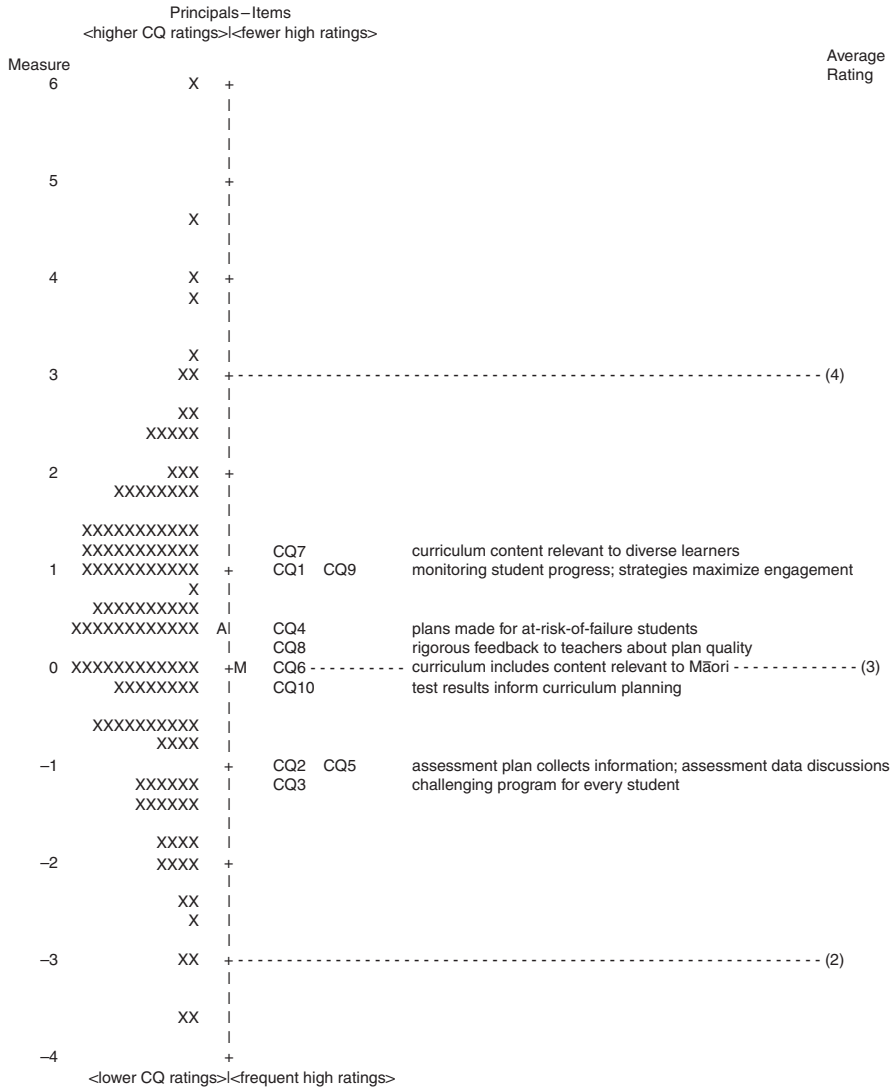
For each of the eight ELP scales the model generates an estimate of the difficulty of achieving a rating of "outstandingly effective" for each item, an estimate of each principal's perceived effectiveness across the set of items and an estimate of the threshold difficulty of responding in the successively higher level response categories (with five categories there are four such estimates). These estimates are reported in a logit metric (Ludlow and Haley, 1995; Wright and Masters, 1982).

Higher rated principals (many "outstandingly effective" ratings) have positive-valued effectiveness estimates; lower-rated principals (fewer "outstandingly effective" ratings) have negative-valued estimates. Harder practice items (fewer "outstandingly effective" ratings) have positive difficulty estimates while easier items (many "outstandingly effective" ratings) have negative estimates. The four threshold estimates within each scale are expected to show a steady increase in their level of difficulty. As shown below, the principal effectiveness and item difficulty estimates simultaneously portray the progressive difficulty of the practices on the CQ scale and the location of each principal on the CQ scale continuum. The CQ scale was selected as the prototype for illustrating the analysis process and the subsequent results. The WINSTEPS software package was used for the analyses (Linacre, 2012).

Variable map

Figure 1 contains the "variable map" for the CQ scale for the principals at Time 1. Variable maps graphically portray the operational definition of the construct that is being measured; in our case there are eight variable maps (only one is presented in this paper). These maps represent one of the key strengths of a Rasch measurement approach to instrument development. That is, if the empirically determined item locations on the map correspond to the a priori formulation of what the scale was intended to measure, then we have strong construct validity evidence for interpreting an individual's score on the scale. Furthermore, this graphical representation enables change score analyses that provide a rich qualitative description of what it means for a person who experiences an intervention and has subsequently moved either higher or lower on the scale (see, e.g. Rollison *et al.*, 2012).

The left-most column contains the logit values corresponding to both the principal effectiveness and item difficulty estimates. The right-most column provides a frame of reference in terms of average rating levels, e.g. where an average category effectiveness



Note: Each "X" represents one principal

Figure 1.
Variable map of curriculum quality, all principals, Time 1

rating of "3" falls. The items are ordered from easiest to rate as "outstandingly effective" (bottom of the map) to hardest to rate "outstandingly effective" (top of the map). To the left, the principals are ordered from lowest overall ratings (bottom of the map) to highest ratings (top of the map). The "A" to the left of the line represents the location of the average CQ effectiveness rating for the principals (their average rating was 30.8 or a Rasch effectiveness logit measure of 0.39). The "M" represents the average item difficulty set to zero – this item centering procedure solves the indeterminacy in scale and metric that results from trying to locate both people and items on a common continuum (De Ayala, 2009; Wright and Panchapakesan, 1969).

Starting at the bottom of the CQ scale it is easiest to rate as "outstandingly effective" items CQ3 (challenging program for every student), CQ2 (assessment plan collects information) and CQ5 (assessment data discussions). These three practices are followed by a ladder-like progression up the scale with slightly harder practices: CQ10 (test results inform curriculum planning), CQ6 (curriculum includes content relevant to Māori), CQ8 (rigorous feedback to teachers about plan quality) and CQ4 (plans made for at-risk-of-failure students). These practices are followed by even harder ones: CQ1 (monitoring student progress) and CQ9 (strategies maximize engagement). Finally, at the top of Figure 1 item CQ7 (curriculum content relevant to diverse learners) defines the highest level and hardest practice to rate as "outstandingly effective" on the CQ variable.

Increasingly harder practices related to CQ are described by the ordering of the items from the bottom of the scale to the top. In particular, the three lower level practices capture relatively routine general tasks associated with data gathering and curriculum purposes. The central level practices address the use of the data to inform various curriculum planning and implementation practices. The highest level of practice reflects the more rigorous tasks of monitoring student outcomes and catering to diversity. The location of the average principal effectiveness rating "A" means that the average principal is perceived as highly effective on the relatively routine tasks of data collection, is satisfactory at curriculum planning and use of test results, and is perceived as minimally effective on higher level monitoring and quality assurance practices. It is consistent with our Rasch measurement expectations that proceeding up the CQ scale means principals engage in increasingly rigorous curriculum planning, implementation and monitoring practices.

Table I is particularly useful because it allows us to find how any one principal's total rating on the CQ scale translates into an "effectiveness" level on the variable map. For example, if a principal had a rating score of 29, the Rasch estimated effectiveness measure would be −0.28 and the principal would be represented as one of the "X" marks adjacent to the location of item CQ10 on the variable map. As noted earlier, the power of the variable map lies in its capability of graphically representing what an individual's score means on any of the eight ELP scales at any particular point in time. A narrative description of what the principal is perceived as doing well can be generated and professional development can be planned as a systematic program of opportunities to master successively higher level practices.

Goodness-of-fit

Rasch model goodness-of-fit analyses generally focus on various statistical or graphical summaries of residuals – the differences between the observed responses that participants provided and the responses expected, i.e. predicted, by the model (Wright and Masters, 1982). The statistical indices may represent a mean squared unstandardized summary (the so-called WINSTEPS generated "mean square outfit"), a mean squared unstandardized, variance weighted summary (mean square infit) or their standardized analogues – the "outfit zstd" and "infit zstd," respectively. Although these indices have different purposes, they tend to be highly correlated and support one another when an item or person demonstrates highly unexpected response variation. These indices have no known distributional form that leads to unequivocal statements about probability levels. Hence, many rules of thumb have been offered (Smith, 1991; Smith *et al.*, 1998). The authors typically use a flexible criterion of +1.4 with the mean squared "infit" to flag unexpected ratings (i.e. a low scoring principal responded much

Score	Measure	SE
10	−8.23E	1.87
11	−6.92	1.07
12	−6.07	0.81
13	−5.50	0.71
14	−5.04	0.65
15	−4.64	0.62
16	−4.27	0.59
17	−3.93	0.58
18	−3.60	0.57
19	−3.28	0.56
20	−2.97	0.55
21	−2.67	0.55
22	−2.37	0.55
23	−2.07	0.55
24	−1.77	0.55
25	−1.47	0.55
26	−1.17	0.54
27	−0.87	0.54
28	−0.58	0.54
29	−0.28	0.54
30	0.01	0.54
31	0.31	0.54
32	0.61	0.54
33	0.90	0.54
34	1.20	0.54
35	1.49	0.54
36	1.79	0.54
37	2.08	0.54
38	2.38	0.54
39	2.67	0.55
40	2.67	0.55
41	2.97	0.56
42	3.28	0.56
43	3.59	0.58
44	3.92	0.59
45	4.26	0.61
46	5.02	0.65
47	5.47	0.70
48	6.03	0.81
49	6.88	1.07
50	8.18E	1.86

Notes: "Score" refers to summative raw score on the curriculum quality scale. "Measure" refers to logit transformation of the raw score into estimate of principal's curriculum quality effectiveness. "SE" refers to the standard error of the logit estimate

Table I.
Curriculum quality score equivalence table

higher than expected on an item or a high scoring principal scored much lower than expected on an item). Our criterion is set relatively low in order to avoid missing surprising ratings that might suggest problems with the items or confusion on the part of the participants. Mean squares less than 1.0 and "zstd" values less than zero represent variation that is less than expected and for Likert-type items this typically can be attributed to relatively more frequent use of the middle level categories.

The CQ fit analysis revealed three interesting items. CQ7 (curriculum content relevant to diverse learners) was the hardest item on the CQ scale. A review of the largest standardized residuals (not shown) revealed that the misfit came from some otherwise higher rated principals who said they were only "minimally" or "satisfactorily effective" in addressing the needs of diverse learners. CQ1 (monitoring student progress) and CQ9 (strategies maximize engagement), in contrast, drew some unexpectedly "highly" and "outstandingly effective" responses from some otherwise lower rated principals who apparently are engaged in some higher level, relatively difficult strategizing practices. These patterns of unexpected responses do not invalidate the meaning of the QC variable; rather, they highlight interesting and specific under- and over-achieving practices within a few schools. Table II contains the traditional reliability and Rasch goodness-of-fit results for all eight scales. In the left-most column are the Cronbach α internal consistency estimates associated with each scale. As is typical of multi-scale inventories such as the ELP, the reliability increases as scale length increases. They are all above 0.80 with an average of 0.88. In addition, there were no negative or zero-valued item-total discrimination correlations within any of the eight scales.

The items within each scale are presented according to their difficulty level. This is a useful strategy in order to check on the relation between item difficulty and item misfit. Of the 14 items out of 82 with a mean square infit > 1.4, eight were the hardest items in their respective scales (including the three hardest in the CQ scale). This is a typical finding when performance, or effectiveness, is assessed. That is, in a situation where an item or task is relatively difficult to accomplish well, there are some people who do not score high on a scale but who are unexpectedly successful on a difficult item because of their unique background and experience. Similar to the analysis of CQ reported above, a review of the largest standardized residuals on each of these eight items found a few otherwise low scoring principals who rated themselves particularly high on these challenging items. This type of misfit is understandable and is useful for measurement purposes because it highlights a strength for some principals that might not otherwise be apparent by just looking at the principal's scale score. Finally, when we consider the seven scales other than CQ (because it was discussed above), there are 11 misfitting items – an average of just 1.4 items per scale which compares favorably with simple statistical chance of at least one misfitting item per scale.

Response options

From a measurement and instrument development perspective, the ideal situation for the five response options is that each is the most probable response at different levels of principal effectiveness. That is, principals with lower levels of effectiveness should score in the lower response categories, more highly effective principals should score in the higher level response categories. This ideal expected pattern is seen in Figure 2. These probability curves for the CQ scale show that each response category is prominently represented as the most probable response for every possible combination of differences in the principal and item estimates. For example, if a principal has an effectiveness estimate that is equal to the difficulty of an item, that difference is zero and the most probable response for the principal on that item is a "3."

The eight sets of threshold estimates are presented in Table III. These estimates correspond to where the curves in Figure 2 intersect. First, it is readily apparent that the threshold estimates display the desired increase in their difficulty order. Second, the estimates themselves are similar across the scales – this means their probability curves

Scale	Item	Item Difficulty	Infit	Outfit	zstd
Goal setting (GS) $\alpha = 0.88$	GS3 – the school's strategic/long-term goals are important to Maori students and their whanau?	1.29	1.49	1.49	3.6
	GS11 – challenging (stretch) learning goals are set for each student?	1.12	0.93	0.93	−0.6
	GS7 – there are clear school-wide targets for the academic achievement of Maori students?	0.94	1.53	1.55	3.9
	GS6 – all the staff are fully aware of the targets in the school's annual plan that are relevant to their area of responsibility?	0.29	1.26	1.26	2.1
	GS10 – everyone is expected to teach in ways that ensure that students at risk of academic failure catch up?	−0.9	0.77	0.77	−2.0
	GS2 – the school's strategic/long-term goals are communicated in clear, concrete terms?	−0.26	1.02	1.02	0.2
	GS9 – everyone has high expectations for the learning of all their students?	−0.37	1.01	1.01	0.1
	GS8 – there is honest non-blaming evaluation of progress toward school targets for student learning?	−0.44	0.72	0.71	−2.6
	GS1 – the school's strategic/long-term goals promote high standards and expectations for all students?	−0.45	0.72	0.71	−2.5
	GS4 – school targets are based on information about what students currently know and are able to do?	−0.97	0.74	0.73	−2.4
	GS5 – school targets promote high standards and expectations for all students?	−0.97	0.72	0.71	−2.6
Strategic resourcing (SR) $\alpha = 0.83$	SR8 – resources are allocated to support the development of school-home partnerships that serve student learning?	1.69	1.10	1.1	0.9
	SR1 – effective teaching resources aligned to school goals are readily available?	0.76	1.13	1.1	1.1
	SR7 – the expertise of families/community is used in ways that serve the school's priority learning goals?	0.74	1.57	1.58	4.2
	SR3 – the timetable reflects the school's priorities for teaching and learning?	−0.35	0.70	0.68	−2.9
	SR2 – there is ready access to teaching and learning resources that engage students at risk of failure?	−0.50	0.90	0.93	−0.9
	SR6 – there is ready access to teaching and learning resources that engage Maori students?	−0.55	1.01	1.03	0.1
	SR5 – students at risk of failure get additional high-quality opportunities to learn?	−0.70	0.67	0.67	−3.2
	SR4 – school routines maximize all students' opportunities to learn?	−1.09	0.86	0.84	−1.2
Curriculum quality (CQ) $\alpha = 0.91$	CQ7 – curriculum in all learning areas includes content relevant to diverse learners?	1.15	1.51	1.50	3.7
	CQ1 – systematic monitoring of each student's progress occurs?	0.99	1.43	1.42	3.2
	CQ9 – strategies are used that maximize the engagement of all students in all classes?	0.94	1.45	1.46	3.3
	CQ4 – students at risk of failure are identified early and plans made to accelerate their progress?	0.40	0.70	0.70	−2.8

(*continued*)

Table II.
Scale and item statistics

Scale	Item	Item Difficulty			
		Difficulty	Infit	Outfit	zstd
	CQ8 – rigorous feedback is given to teachers about the quality of their schemes/unit plans?	0.16	0.68	0.67	−3.0
	CQ6 – curriculum in all learning areas includes content relevant to the identity of Maori students?	−0.01	0.73	0.72	−2.5
	CQ10 – there is routine discussion of the results of common tests or tasks in teaching teams, and staff use these discussions to inform their curriculum planning?	−0.30	0.81	0.81	−1.7
	CQ5 – discussions of student assessment data focus on the relationship between what was taught and what students learned?	−1.0	1.13	1.13	1.1
	CQ2 – there is a school/departmental assessment plan to collect the information needed to monitor progress on priority learning goals?	−1.09	0.7	0.69	−2.8
	CQ3 – every student experiences a challenging program?	−1.23	0.83	0.83	−1.4
Quality of teaching (QT) $\alpha = 0.92$	QT8 – any teaching problems are discussed with a colleague with relevant expertise?	2.57	1.83	1.92	5.6
	QT5 – appraisal focusses on improving teaching practice and student outcomes?	0.38	0.76	0.76	−2.1
	QT4 – early identification and support is provided for teachers who are having difficulty helping students reach important academic and social goals?	0.31	0.68	0.68	−3.0
	QT9 – mandated procedures such as attestation and appraisal are used as serious opportunities for the improvement of teaching?	0.07	0.79	0.79	−1.8
	QT6 – assessment data are used to improve teaching?	−0.25	1.16	1.16	1.4
	QT1 – everybody shares the responsibility for students' academic and social learning?	−0.30	1.04	1.04	0.3
	QT7 – students provide feedback to teachers on the effectiveness of their teaching?	−0.69	0.98	0.98	−0.1
	QT2 – those with particular expertise are used to help other teachers in the school to develop their knowledge and skills?	−0.74	0.89	0.91	−0.9
	QT3 – there is challenge and support to improve teaching for those teachers whose students remain disengaged?	−1.36	0.8	0.78	−1.7
Collaborative teacher learning and development (CT) $\alpha = 0.86$	CT9 – professional development opportunities enable teachers to develop the knowledge and skills necessary to provide quality teaching for diverse learners?	1.28	1.58	1.58	4.2
	CT6 – decisions to maintain or to change particular teaching approaches are based on evidence about their impact on students?	0.81	0.96	0.96	−0.3
	CT8 – professional development opportunities enable teachers to develop the knowledge and skills necessary to provide quality teaching for Maori learners?	0.51	0.92	0.92	−0.7

Table II.

(*continued*)

Scale	Item	Item Difficulty	Infit	Outfit	zstd
	CT4 – staff meetings include serious discussions about how to improve teaching and learning?	0.35	0.69	0.69	−3.0
	CT1 – student achievement patterns are analyzed and used to plan professional learning priorities?	0.01	1.13	1.12	1.1
	CT7 – a range of evidence sources is used by teachers to evaluate the effectiveness of their teaching?	−0.26	0.93	0.93	−0.6
	CT10 – professional development and learning is evaluated in terms of its impact on students?	−0.44	0.93	0.92	−0.6
	CT3 – adequate opportunities are provided for teachers to discuss why they might need to change their practice?	−0.48	0.83	0.87	−1.5
	CT5 – systematic opportunities are provided for teachers to improve their teaching through observing the teaching of effective colleagues?	−0.87	1.18	1.17	1.5
	CT2 – there is open discussion of students' results, and teachers help each other develop more effective teaching strategies?	−0.91	0.79	0.85	−1.8
Safe and orderly environment (SO) $\alpha = 0.88$	SO1 – staff work in a safe, supportive and orderly environment?	1.69	1.46	1.52	3.5
	SO8 – the school is a positive environment in which student learning is the central focus?	1.13	1.39	1.45	3.0
	SO10 – the school is a positive environment for everyone, whatever their culture?	0.87	1.72	1.73	5.0
	SO7 – student views about the school culture and how to improve it are taken seriously?	−0.08	0.77	0.80	−2.0
	SO5 – there is a consistent school-wide approach to student behavior management?	−0.37	0.58	0.58	−4.0
	SO6 – timely support with student behavior issues is given to staff?	−0.44	0.94	0.95	−0.4
	SO4 – problems between teachers and parents are resolved in a fair and timely way?	−0.47	0.58	0.58	−4.0
	SO3 – problems between teachers and students are resolved in a fair and timely way?	−0.59	0.72	0.74	−2.4
	SO2 – staff views about the school culture and how to improve it are taken seriously?	−0.66	0.78	0.79	−1.9
	SO9 – there is regular monitoring of the extent to which students feel safe at school?	−1.08	0.87	0.85	−1.1
Families and community (FC) $\alpha = 0.82$	FC7 – accurate information about school academic and social learning performance is available to the community?	1.35	1.94	1.93	6.4
	FC4 – staff are responsive to families' views about their child's learning needs?	0.53	0.92	0.92	−0.6
	FC2 – the school provides parents with opportunities to learn how to support their child's school learning?	0.48	0.90	0.90	−0.8
	FC3 – parents understand the achievement levels of their children in relation to national benchmarks?	0.16	0.72	0.73	−2.6
	FC8 – school/community relations are focussed on enhancing educational outcomes for students?	−0.21	0.66	0.66	−3.2

(*continued*)

Table II.

Scale	Item	Item Difficulty			
		Difficulty	Infit	Outfit	zstd
	FC6 – the school works in partnership with local Maori leaders to support Maori aspirations?	−0.30	1.05	1.04	0.5
	FC5 – there are systematic processes for gaining parent and community feedback about the school?	−0.38	0.61	0.62	−3.3
	FC1 – class programs are discussed with parents so that parents understand what their child is being taught?	−1.63	1.11	1.15	1.0
Principal leadership (PL) $\alpha = 0.93$	PL4 – leading useful discussions about the improvement of teaching and learning?	1.07	1.05	1.06	0.5
	PL11 – earning the respect of the different ethnic communities served by the school?	0.88	1.12	1.15	1.0
	PL5 – identifying and resolving conflict quickly and fairly?	0.87	1.09	1.08	0.8
	PL1 – using research on teaching and learning to inform important school decisions?	0.84	1.12	1.09	1.0
	PL9 – earning the respect of all of the staff?	0.74	0.97	0.97	−0.2
	PL2 – learning alongside teachers about how to improve teaching and learning?	0.37	1.47	1.47	3.6
	PL10 – earning the respect of the wider community?	0.08	0.85	0.80	−1.3
	PL14 – saying what s/he thinks and explaining why?	0.02	0.95	0.95	−0.4
	PL12 – seeking high-quality information about the situation before making a final decision?	−0.14	0.96	0.93	−0.3
	PL16 – making tough decisions when necessary?	−0.20	1.36	1.33	2.7
	PL3 – serving the interests of the whole school rather than of particular interest groups?	−0.41	0.82	0.81	−1.5
	PL15 – actively seeking others' views?	−0.5	0.9	0.88	−0.8
	PL7 – maintaining integrity in difficult situations?	−0.57	0.92	0.89	−0.6
	PL8 – showing both personal and professional respect for staff?	−0.91	0.69	0.67	−2.7
	PL6 – promoting and modeling the values of this school?	−1.04	0.81	0.77	−1.6
	PL13 – being open to learning and admitting mistakes?	−1.1	0.84	0.82	−1.2

Table II.

all resemble Figure 2. Third, the separation and spread in the threshold estimates is excellent. These findings mean that the response categories are understood as clearly distinct levels, they are being used the same way regardless of scale, and that the rating scale model was an appropriate choice rather than a partial credit model.

Dimensionality checks

Rasch's assertion of "uniformity of content" is typically re-expressed as "unidimensionality is defined as the existence of one latent trait underlying the data" (Hattie, 1984, p. 139). This characteristic of the items is critical since it lays the foundation for the likelihood-based procedures through which the Rasch parameters are estimated (Stout, 1987). The statistical problem is that there is "no universally accepted technique or set of rules to determine the number of factors to retain when assessing the dimensionality of item response data" (Slocum, 2005, p. 3). Two frequently cited factor analysis criteria for establishing unidimensionality include Reckase's (1979) suggestion that "for acceptable calibration, the first factor should

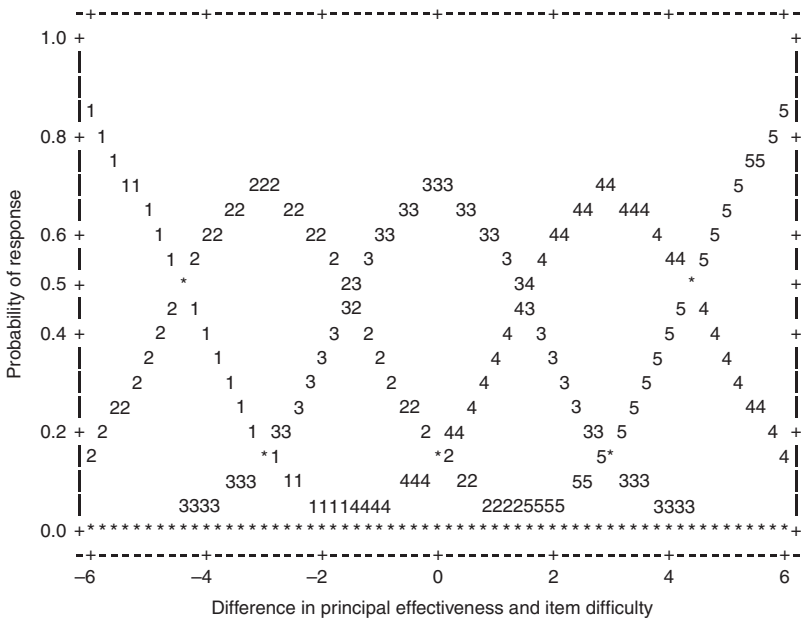


Figure 2.
Curriculum quality response probabilities (category characteristic curves)

	Educational leadership practices scales							
Category	Curriculum quality (CQ)	Collaborative teacher learning and development (CT)	Families and community (FC)	Goal setting (GS)	Principal leadership (PL)	Quality of teaching (QT)	Safe and orderly environment (SO)	Strategic resourcing (SR)
1								
2	−4.43	−4.09	−3.45	−4.29	−5.61	−4.34	−3.96	−4.43
3	−1.47	−1.41	−1.39	−1.62	−2.22	−1.74	−1.52	−1.46
4	1.5	1.29	1.06	1.36	1.83	1.33	0.92	1.22
5	4.39	4.21	3.78	4.55	6.0	4.74	4.57	4.67

Note: Threshold estimates refer to the estimates of difficulty of responding in suggestively higher categories

Table III.
Threshold estimates

account for at least 20 percent of the test variance" (p. 228) and the suggestion that a 3-to-1 ratio of the magnitude of the first eigenvalue to the second eigenvalue "constitutes a dominant first factor" (Reise and Revicki, 2015, p. 18). Stout (1987), in addition, proposed similar criteria for establishing "essentially unidimensional" (p. 597).

Three forms of unidimensionality checks were performed: factor analyses of the raw data (Reckase, 1979), principal component analysis of the Rasch residuals (Ludlow, 1983) and parallel analyses of simulated data (O'Connor, 2000). For the raw data we seek first factors that account for more than 20 percent of the variance, first-to-second eigenvalue ratios that exceed three and Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy indices > 0.7 (Kaiser, 1970; Kaiser and Rice, 1974). For the residuals we seek first-to-second eigenvalue ratios that are near 1, plots of the first two components that resemble circular patterns and KMO values near 0. The parallel

analysis will establish the magnitudes of eigenvalues typically extracted from random data based on equivalent numbers of items.

A principal axis factor analysis was performed on the raw data Pearson correlation matrices for each of the eight scales. Principal component analyses were performed on the residual Pearson correlation matrices for each scale. The distinction between the two procedures is how measurement error is treated. With factor analysis an estimate of the shared variance is included on the diagonal of the correlation matrix – error is not explicitly included in the determination of common factors. This means each scale is analyzed from the perspective of just the covariance the items share. With principal components there is no distinction between common and error variance. The residuals from the Rasch model are assumed to consist of nothing but error variance – there should be no common variance. Parallel eigenvalue analyses were then performed based on $n = 148$, 100 simulations, and from eight to 16 items. All analyses employed SPSS (Version 22). The results are presented in Table IV.

For the raw data, seven of the eight ELP scales have KMOs in the "meritorious" or "marvelous" levels of 0.8 or 0.9, respectively. The other scale is acceptable at 0.77. Their average is 0.85. For the residuals, all KMOs are < 0.05 – demonstrating that the Rasch model has accounted for sufficient variability in estimating the parameters and that, consequently, there is insufficient residual variability suitable for factoring.

The ratios of the first-to-second eigenvalues for the raw data all exceed the 3-to-1 suggestion of Reise and Revicki (2015). Their average is 4.1. The ratios for the residuals are all near 1, their average is 1.3. These results from the residuals are mirrored in the various parallel analyses of random data where the first two eigenvalues all show ratios near 1 (their average was 1.11). In addition, the percent of variance for the first factors from the raw data are all > 20 percent, seven of the eight are twice that amount.

Finally, the first two unrotated factors for the raw data (Embretson and Reise, 2000) all resemble the pattern in Figure 3 for CQ – the items load high on a first factor and lower on a second factor. In contrast, the first two unrotated principal components in Figure 4 for the CQ residual matrix resembles the circular pattern characteristic of a random pattern relationship among the items (Ludlow, 1983, 1985, 1986). This pattern was also seen in the other seven ELP scales. Taken together, the reliability analyses, factor analyses, principal component analyses and parallel analyses all provide strong evidence that each of the eight ELP scales is an "essentially unidimensional" construct.

Scale invariance

We turn now to the question of whether the eight ELP scales retain the same meaning at both time points. That is, are the scales invariant in their item locations for the principals? For these next analyses we highlight the PL scale. Figure 5 contains the PL variable maps for the principals at Time 1 and Time 2. Briefly, at Time 1 the items move upwards from relatively routine leadership practices such as PL15 (actively seeking others' views), PL13 (open to learning and admitting mistakes) and PL8 (showing staff respect) to harder more rigorous leadership practices such as PL1 (using research to inform school decisions), PL5 (resolving conflict quickly and fairly) and PL4 (leading useful discussions about improvements). This ordered progression is again consistent with the scale expectations presented earlier.

The measurement issue here is the extent to which the understanding and meaning of the PL scale has stayed consistent (invariant) from Time 1 to Time 2 (Figure 5). If the item ordering and pairs of difficulty estimates are the same (within their standard errors) then we have evidence of scale invariance. This means that the same rating score at two

Category	Curriculum quality (CQ) scale		Collaborative teacher learning and development (CT) scale		Families and community (FC) scale		Goal setting (GS) scale		Principal leadership (PL) scale		Quality of teaching (QT) scale		Safe and orderly environment (SO) scale		Strategic resourcing (SR) scale		Average
	O[a]	Z[a]	O	Z	O	Z	O	Z	O	Z	O	Z	O	Z	O	Z	
$O\Lambda_1$[b]	3.9	1.9	4.4	1.7	3.3	1.7	5.1	2.1	7.8	2.7	4.0	2.1	4.9	2.1	3.7	1.8	
%S[c]	39.4		44.2		42.5		46.0		48.6		44.8		49.3		45.7		45.1
$O\Lambda_2$[d]	1.3	1.5	1.1	1.4	1.0	1.5	1.1	1.7	1.5	2.0	1.1	1.5	1.0	1.8	1.0	1.5	
$O\Lambda_1/O\Lambda_2$[e]	3.0	1.3	4.0	1.2	3.4	1.1	4.6	1.2	5.2	1.4	3.6	1.4	4.9	1.2	3.7	1.2	(O = 4.1, Z = 1.3)
KMO[f]	0.83	0.04	0.90	0.02	0.77	0.04	0.87	0.03	0.92	0.03	0.84	0.04	0.85	0.04	0.84	0.04	(O = 0.85, Z = 0.04)
$P\Lambda_1$[g]	1.43		1.43		1.38		1.45		1.60		1.40		1.43		1.38		
$P\Lambda_2$[h]	1.30		1.30		1.22		1.31		1.47		1.25		1.30		1.22		
$P\Lambda_1/P\Lambda_2$[i]	1.10		1.10		1.13		1.12		1.09		1.12		1.10		1.13		1.11

Notes: [a]O = Observed data, Z = residual data; [b]$O\Lambda1$ = first observed eigenvalue; [c]%S2 = first factor variance; [d]$O\Lambda2$ = second observed eigenvalue; [e]$O\Lambda1/O\Lambda2$ = ratio of first to second observed eigenvalues; [f]KMO = Kaiser-Meyer-Olkin measure of sampling adequacy; [g]$P\Lambda1$ = first parallel analysis eigenvalue; [h]$P\Lambda2$ = second parallel analysis eigenvalue; [i]$P\Lambda1/P\Lambda2$ I = ratio of first to second parallel analysis eigenvalues

Table IV.
Eigenvalue analysis

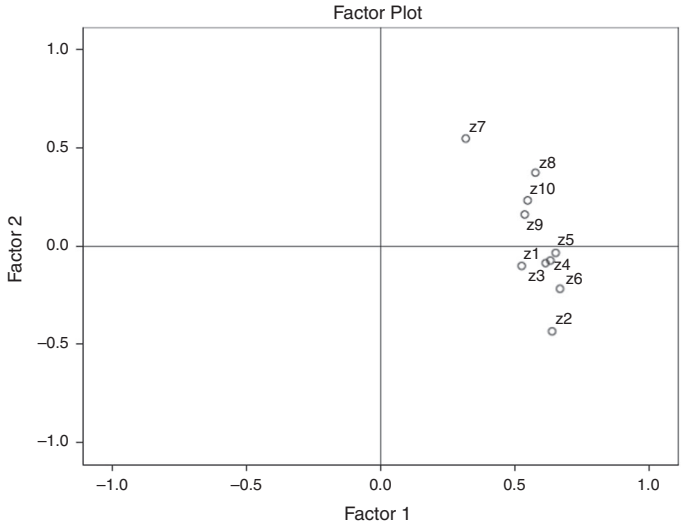


Figure 3.
Unrotated factor analysis pattern for curriculum quality

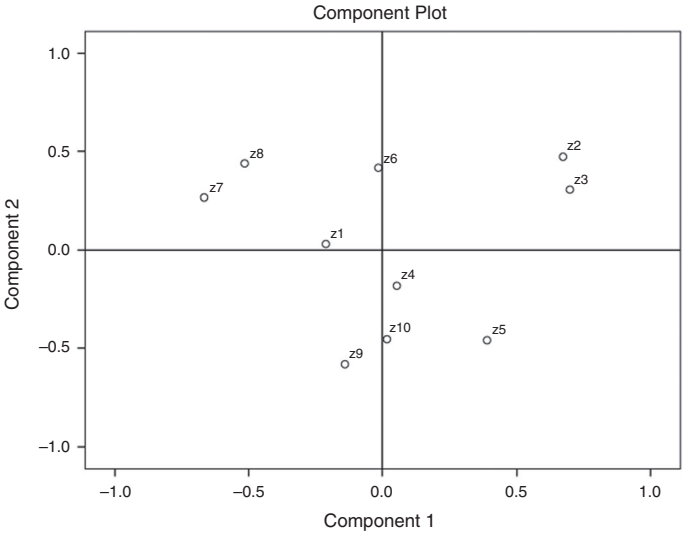
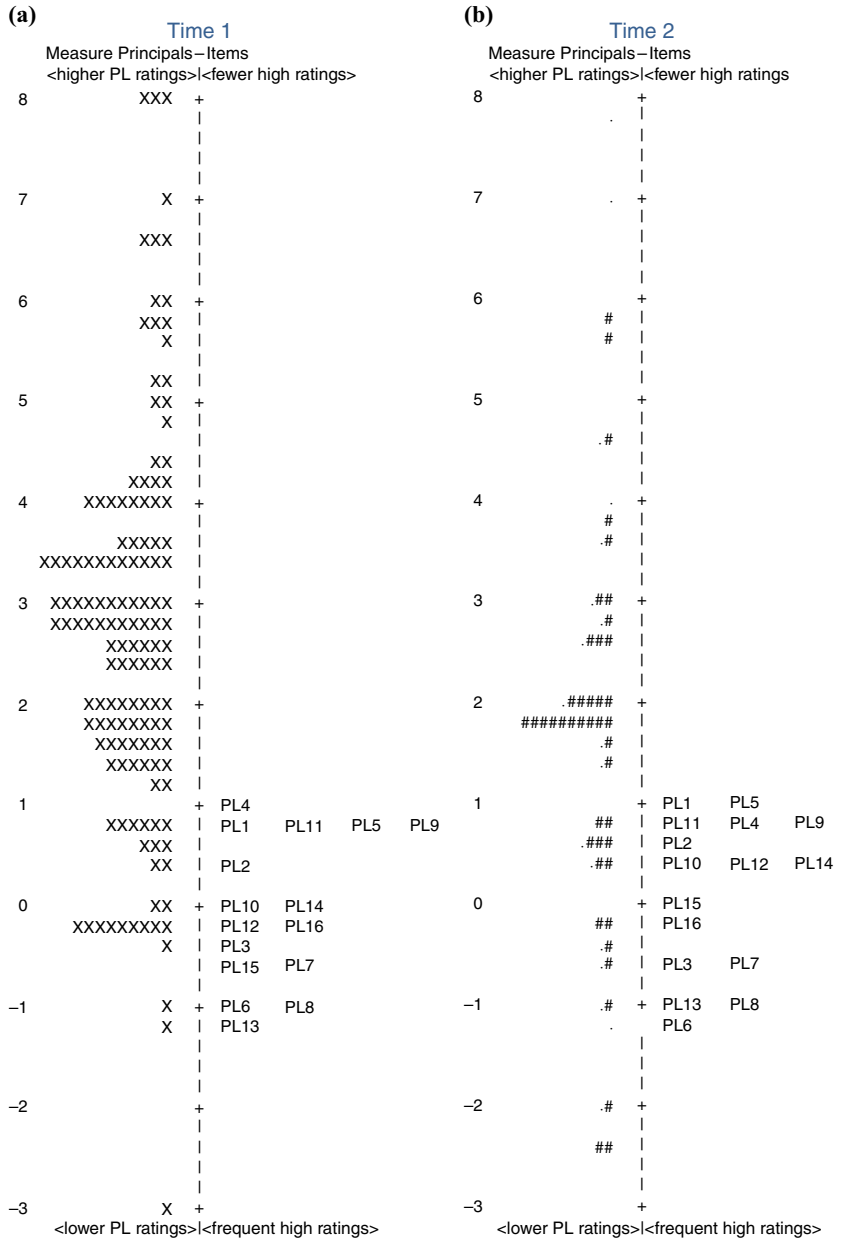


Figure 4.
Unrotated principal component pattern for curriculum quality

different times will have the same meaning and will, theoretically, have been generated by the same pattern of lower, mid-level and higher level ratings. This property of invariance over time, if achieved, would mean that the scale is suitable for use in analyses of change – such as evaluations of professional development interventions.

It is important to note that a sample size of 148, while statistically sufficient for estimating Rasch model parameters (De Ayala, 2009; Wright and Masters, 1982), will lead to standard errors for these estimates that will be larger than if a larger n had been available. This means that items located adjacent to one another will have confidence intervals around



Notes: Each "X" is one principal; each "#" is 2 principals; each "." is 1 principal

Figure 5.
Variable maps of principal leadership items, Time 1 and Time 2, all principals

them that may overlap (meaning their "true" location could be either above or below their current position) and their item locations may shift from Time 1 to Time 2 simply due to measurement error. The procedures employed below to assess invariance do take into account the measurement precision represented by the standard errors.

A close inspection of Figure 5 for Time 1 and Time 2 reveals a similar order to the item difficulty estimates, although the Time 1 estimates are slightly more clustered while there is a slightly more uniform spread in the Time 2 estimates. A scatterplot of the ten pairs of item estimates (not shown) from Time 1 and Time 2 ($r = 0.957$, $p < 0.001$) revealed that none of the paired estimates fell outside the 95 percent confidence interval bands. The magnitude of the correlation is also evidence of the strong test-retest, stability-over-time, reliability of the estimates. In addition, z-tests (Wright and Stone, 1979) revealed no statistically significant differences between the paired item difficulty estimates for the principals at Time 1 and Time 2.

These analyses were performed for the other seven ELP scales at Time 1 and Time 2. Out of a total 82 item comparisons there were only 11 item difficulty differences – six showed statistically significant differences where principals rated themselves more conservatively (harder) at Time 2 than Time 1 while five items were less challenging at Time 2 than Time 1. The test-retest reliabilities ranged from 0.952 (CQ) to 0.991 (SR).

The next analyses addressed the invariance of the eight ELP scales as perceived by the teachers at both times and the teachers vs the principals at both times. If invariance across time and educator can be demonstrated in these analyses, it is then possible to form a single set of ELP scales that are appropriate for both principals and teachers, and measure change from Time 1 to Time 2 on a common equated scale. The following analyses were performed on the PL scale in order to maintain consistency with the principals' results.

The pairs of item estimates for the teachers on the PL scale at both times were plotted and their test-retest, stability-over-time reliability was $r = 0.993$ ($p < 0.001$). The z-tests were conducted and item PL10 (earning the respect of the wider community) was slightly easier at Time 2 for the teachers to rate as "outstanding effective." These analyses were performed for the other seven ELP scales at both times and, of 82 comparisons there were only 15 item difficulty differences – five showed statistically significant differences where teachers perceived the items as less challenging for principals at Time 1 than Time 2 while for ten items they perceived them as less challenging for principals at Time 2 than Time 1. The test-retest reliabilities ranged from 0.978 (CQ) to 0.993 (PL).

This number of significant differences amounts to 1.4 items (for principals) and 1.8 items (for teachers) per scale being rated differently at Time 2. If these 82 tests were statistically independent, then our total number of expected significant differences due to chance would be $82 \times 0.05 = 4.1$ items at the $\alpha = 0.05$ level. Given that the cross-scale comparisons are correlated, since the same principals and many of the same teachers responded to all the scales, a Bonferroni adjustment set the α at $0.05/8 = 0.006$ per comparison at which point there were only two items that showed this magnitude of change for the principals and four such items for the teachers. Overall, this high degree of similarity in the item estimates suggests that both principals and teachers were, as anticipated, consistent in their understanding of the eight ELP scales at both Time 1 and Time 2.

Comparing teachers' and principals' ratings of PL

The next analyses compared the teachers' ratings vs the principals' at both times on the PL scale. Formally, these are referred to as differential item function analyses (DIF). A plot of the paired item estimates for the teachers and principals at Time 1 revealed obvious differences in the estimates, the correlation between the estimates was only $r = 0.42$ ($p = 0.11$), and Mantel-Haenszel (MH) χ^2 analyses performed specifically to test

for DIF confirm that teachers and principals differed on some items in how they rated the effectiveness of the principals' leadership practices.

On the one hand, the principals rated themselves less effectively than did the teachers on PL1 (using research to inform school decisions), PL4 (leading discussions about improvements) and PL11 (earning respect of ethnic communities) – a result that may be attributed to teachers knowing less about these practices than the principals. On the other hand, the principals rated themselves more effectively than the teachers on PL3 (serving interests of the whole school rather than particular groups), PL13 (open to learning and admitting mistakes) and PL15 (actively seeking others views). These practices, in contrast to the previous set, are those that teachers are more likely to have had personal experience of and are, perhaps, ones on which principals and teachers are in some disagreement. One implication of these differences is the need for greater transparency in school settings about the leadership practices required to promote improvement in teaching and learning. An additional implication from a research perspective is the need to include attention to the cause of variation across time to establish whether changes in ratings of principals are attributable to actual changes in their practice, or to changes in teachers' understanding of practices described in survey items, or of the rating scale itself.

Since the above results suggest there may be differences in how principals and teachers understand the job that principals actually do, the next analysis of the principals' and teachers' ratings at Time 2 was expected to show more mutual understanding of the ELP practices, in general, and PL, in particular because they had spent another six months together. The teacher vs principal item difficulty estimate plot was generated ($r = 0.51$, $p < 0.05$), z-tests, and MH analyses were performed and items PL13 (open to learning and admitting mistakes) and PL15 (actively seeking others views) showed meaningful shifts in how the principals perceived themselves relative to the teachers. In both cases, the principals rated themselves lower than their Time 1 ratings and more consistently with how their teachers rated them.

These various forms of principal vs teacher invariance analyses were performed for each of the other seven ELP scales. Differences between the two educator groups tended to be greatest at Time 1 while the magnitude and direction of those differences often diminished at Time 2.

In summary, the series of analyses performed upon the portfolio of ELP scales suggest that these scales possess the critical measurement property of invariance across time and educator group. This means principals and teachers have essentially the same understanding of the levels of difficulty differentiating these leadership practices, and the stability of the scale means it is suitable for the analyses of PL effectiveness changes over time.

The progression of challenge in ELP

The next stage of analysis involved the development of qualitative construct descriptions to capture the essence of the progression in leadership practices across all of the scales. To do this, a schematic version of the variable maps was created for each scale based on both principals' and teachers' data. The schematic version comprised a vertical line for each scale with a notation for the practice locations for principals (on the left of the line) and teachers (on the right of the line). These schematic versions were designed to indicate the relative, rather than exact positioning of each practice for each scale with approximate spacing in the vertical positioning of the practice notations indicating those practices toward the top, middle and lower end of the variable maps.

We compared the positioning of practices in the principal and teacher schematics to establish if there were notable differences in the progression of practices between the two respondent groups. We asked, for example, "Were the practices that principals deemed most difficult to rate as outstandingly effective also deemed similarly difficult by teachers?" Comparison of the sequence between respondent groups revealed that the sequence of practice difficulty was either the same (as in the case of the goal-setting dimension) or only slightly different between principals and teachers (e.g., the item "the expertise of families/community is used in ways that serve the school's priority learning goals" (SR7) was the most difficult practice in the strategic resourcing scale for teachers, and the second most difficult for principals to rate as outstandingly effective). No practices were positioned toward the top of the schematic for one group and toward the bottom for the other. Having established similarities in the broad positioning of practices (all of which were statistically tested and established in the preceding sections), we continued with the second phase of the construct description analyses.

In this phase, a more elaborate version of the schematic map was created which included the full item wordings organized to display all practices in the eight scales. This enabled us to identify the nature of the practices that principals and teachers deemed easiest and hardest to rate as effectively practiced in their school. We were interested in the extent to which our hypotheses about item difficulty within each scale were confirmed and in the progression of item difficulty across all scales.

Leadership practice progressions within scales

To check our hypotheses about the nature of progressions in leadership practice difficulty within each scale we referred to the 3-4 items revealed on the Rasch variable maps to be hardest to rate as outstandingly effective (at the top of the map) and easiest to rate as outstandingly effective (at the bottom of the map). We analyzed those items to establish if they confirmed or disconfirmed out hypothesis for each dimension, or if they revealed another pattern of note. Data from the Time 1 administration to teachers were used for these analyses.

For the goal-setting dimension, we hypothesized that items about setting goals would be easier to rate as outstandingly effective than those that stipulate rigorous inquiry into evidence for the setting and evaluation of goal achievement. Unexpectedly, items about setting goals (rather than rigorous inquiry as part of goal setting and evaluation) were located at both the top and bottom of the variable map indicating some of these items were the easiest and others were the hardest to rate as outstandingly effective[1]. We noted two characteristics of the items that were hardest. The goal setting required a focus on a particular group of learners (Maori) or a personalized approach involving goals for each and every student:

> GS7 – there are clear school-wide targets for the academic achievement of Maori students.
>
> GS3 – the school's strategic/long-term goals are important to Maori students and their whanau.
>
> GS11 – challenging (stretch) learning goals are set for each student.

The items that were easiest were also focussed on goal setting (not rigorous inquiry). One possible reason they were deemed easier than those outlined above is that the

wording "all students" rather than "each student" implies goals that are generally and collectively relevant rather than relevant to each individual:

GS9 – everyone has high expectations for the learning of all their students.

Another reason may be that the easiest items were tightly connected to school accountability mechanisms which require the submission of targets to the Ministry of Education:

GS4 – school targets are based on information about what students currently know .and are able to do

GS5 – school targets promote high standards and expectations for all students.

For the strategic resourcing dimension, we hypothesized that items about resourcing for learners generally would be easier to rate as more highly effective than those about resourcing for groups of learners with particular needs. Unexpectedly, items about resourcing for learners with particular needs were among the easiest:

SR2 – there is ready access to teaching and learning resources that engage students at risk of failure.

SR6 – there is ready access to teaching and learning resources that engage Maori students.

SR5 – students at risk of failure get additional high-quality opportunities to learn.

SR4 – school routines maximize all students' opportunities to learn.

This finding perhaps reflects the positive impact of targeted funding initiatives for those groups at the time of the study. Items that were hardest shared a focus on home school partnerships, either as the goal to be resourced, or as the resource to achieve goals:

SR8 – resources are allocated to support the development of school-home partnerships that serve student learning.

SR7 – the expertise of families/community is used in ways that serve the school's priority learning goals.

Unexpectedly, an item about resourcing for learners generally was also among the hardest:

SR1 – effective teaching resources aligned to school goals are readily available.

For the CQ dimension we hypothesized that items focussed on the more administrative aspects of dealing with CQ (ensuring plans are in place, for example) would be easier to rate as outstandingly effective than those requiring high-quality curricula for all learners in all learning areas. As expected, the hardest items in this scale were those requiring high-quality curricula for all learners in all learning areas – these referred in particular to the relevance of content, the effectiveness of engagement strategies and the monitoring of each student's achievement:

CQ7 – curriculum in all learning areas includes content relevant to diverse learners.

CQ1 – systematic monitoring of each student's progress occurs.

CQ9 – strategies are used that maximize the engagement of all students in all classes.

For the quality teaching dimension we hypothesized that items focussed on practices with the potential to be considered as collegial (about helping and sharing, for example) might be easier to rate as outstandingly effective than those emphasizing more rigorous progression practices (such as improvement, discussion of problems, feedback and challenge). As expected, items that involved addressing teaching problems and that focussed on improvement were rated as harder than those with a more collegial emphasis:

> QT8 – any teaching problems are discussed with a colleague with relevant expertise.
>
> QT4 – early identification and support is provided for teachers who are having difficulty helping students reach important academic and social goals.
>
> QT5 – appraisal focusses on improving teaching practice and student outcomes.

For the teacher development dimension we hypothesized that items requiring engagement with evidence (including data about their own students' progress) might be harder to rate as outstandingly effective than those items without an emphasis on evidence. As expected, among the hardest items was one requiring attention to evidence as the basis for teaching decisions:

> CT6 – decisions to maintain or to change particular teaching approaches are based on evidence about their impact on students.

Additionally, the hardest items in this scale also required a focus on diverse learners and Maori learners in particular:

> CT9 – professional development opportunities enable teachers to develop the knowledge and skills necessary to provide quality teaching for diverse learners.
>
> CT8 – professional development opportunities enable teachers to develop the knowledge and skills necessary to provide quality teaching for diverse learners.

For the safe and orderly environment dimension we hypothesized that items focussed on routine management and monitoring would be easier to rate as outstandingly effective than those involving resolving problems in relation to the school environment. As expected, the items rated easiest were those with a routine management and monitoring orientation. They were about procedural aspects for promoting a safe and orderly environment:

> SO6 – timely support with student behavior issues is given to staff.
>
> SO2 – staff views about the school culture and how to improve it are taken seriously.
>
> SO9 – there is regular monitoring of the extent to which students feel safe at school.

Unexpectedly, the items rated hardest were not those involving resolving problems in relation to the school environment, but those seeking an evaluation of the actual quality of the school environment (being safe, supportive, orderly, positive) rather than the processes in place for promoting it:

> SO1 – staff work in a safe, supportive and orderly environment.
>
> SO8 – the school is a positive environment in which student learning is the central focus?
>
> SO10 – the school is a positive environment for everyone, whatever their culture.

For the dimension about connections with family and community we hypothesized that items requiring the provision of information would be easier to rate as outstandingly effective than those requiring more two-way interactions and genuine partnership between students' parents and school or between settings. As expected, items that position families, schools and communities as partners in children's education were among the hardest items in this scale:

> FC4 – staff are responsive to families' views about their child's learning needs.
>
> FC2 – the school provides parents with opportunities to learn how to support their child's school learning.

Additionally, the hardest item in this scale was one about the accuracy of information schools make available to the community:

> FC7 – accurate information about school academic and social learning performance is available to the community.

This is likely explained by the timing of this survey's administration coinciding with much public criticism about how teachers are reporting to parents and the community about student achievement.

Also as expected, items that focus on the provision of information (rather than genuine partnership) in both the family-school and school-family direction were among the easiest in this scale:

> FC5 – there are systematic processes for gaining parent and community feedback about the school?
>
> FC1 – class programs are discussed with parents so that parents understand what their child is being taught.

The broad progression of leadership practice

To consider the progression of leadership practice difficulty across scales we asked, "What is common to the leadership practice items that were found to be easier and harder to rate as effectively practiced, regardless of the particular dimension they relate to?" For example, two items in different scales that were toward the bottom of the schematic map indicated that compliance practices are relatively easy to perform effectively. One item about CQ ("there is a school/departmental assessment plan to collect the information needed to monitor progress on priority learning goals," CQ2) and a supportive and orderly environment item ("there is regular monitoring of the extent to which students feel safe at school," SO9) both described data collection routines that are now widely established in New Zealand schools.

Practices which school leaders were more effective at described aspirations and practices involving compliance and the provision of opportunities (see the "Item focus" column of Table V). For example, school leadership is relatively effective at aspirational practices including setting goals and targets and promoting high standards and expectations. We have already noted that some of these practices, such as setting and reporting on annual goal and targets, are required by legislation. Similarly, matters of compliance were viewed as relatively effectively carried out – for example, "there is a school/departmental plan to collect the information needed to monitor progress on priority learning goals" (CQ2). The most effective practices were also characterized by wording that emphasized the provision of opportunity, rather than the achievement of a particular outcome. For example, "systematic opportunities are provided for teachers

Table V.
Routine and rigorous leadership practices

	Item focus	Aspects of practice
Routine practices	Aspirations	Targets; goals; expectations
	Compliance	Monitoring; systems; processes; procedures; routines; legislated requirements for goal setting
	Opportunities	Opportunity; access; discussion; planning
Rigorous practices	Specific problems	Resolving conflict; resolving problems
	Diversity	Diverse learners; at risk learners; Māori students; "all students"; "each student's"
	Relevant expertise	Home-school partnerships; engaging the expertise of families/communities
	Improvement	Improvement; change; new knowledge and skills
	School environment	Safe; supportive; orderly; positive

to improve their teaching through observing the teaching of effective colleagues" (CT5). Such an opportunity could be viewed as merely a matter of administrative efficiency – the timetabling of "opportunities" for classroom observation is an easier practice than one that requires that such observation achieves an important change. The same point could be made about practices requiring access ("there is ready access to teaching and learning resources that engage students at risk of failure," SR2) and those that require discussion rather than action ("class programs are discussed with parents so that parents understand what their child is being taught," FC1).

Practices at which school leadership was less effective were focussed on problem solving, diversity, the engagement of those with relevant expertise, improvement-focussed practices and the environment of the school. These aspects of practice, collectively, indicate a leadership construct of rigorous practice that school leadership is likely to be less effective at than the more routine practices discussed above.

Responses indicated that school leadership is less effective with regard to problem-specific practices such as identifying and supporting teachers who are having difficulty and helping students reach important academic and social goals (see QT4) than at practices not involving problems. School leadership is also less effective with regard to responsiveness to diverse learners, as evident in the positioning of the items "professional development opportunities enable teachers to develop the knowledge and skills necessary to provide quality teaching for diverse learners" (CT9), and "curriculum in all learning areas includes content relevant to the identity of Māori students" (CQ6). Items with very similar wording produced very different effectiveness ratings if they referred to particular student groups rather than all students. For example, the more general item ("school targets promote high standards and expectations for all students," GS5) was the practice at which principals perceived they were most highly effective, but when the same item was focussed on Māori students ("there are clear school-wide targets for the academic achievement of Māori students," GS7) it was found to be the least effective practice for principals.

Engaging relevant expertise, including the expertise of parents and communities, was also shown to be more difficult than other leadership practices. This was indicated by the fact that the item "any teaching problems are discussed with a colleague with relevant expertise" (QT8) was rated the least effective practice by both principals and teachers. This indicates a shortage of expertise required to solve the complex and demanding problems teachers face. Similarly, items about professional development enabling the development of new knowledge and skills, and changes to teaching

approaches being based on evidence, and the conduct of serious discussions about the improvement of teaching and learning, all indicate the challenges involved in leading the professional learning of teaching in ways that produce real improvement. Furthermore, items about the positive environment of the school received low effectiveness ratings. Neither principals nor teachers considered leadership effective in creating a school environment that is safe, supportive, orderly or positive. This finding is particularly problematic given that effectiveness on this dimension is probably a necessary though not sufficient condition for effective leadership on the other dimensions (Antoniou, 2013).

Discussion

There is a growing body of evidence about the links between instructional leadership and student outcomes (Goldring, 2009a; Robinson *et al.*, 2008). This study has used a meta-analysis of this research (Robinson *et al.*, 2008), along with some supplementary evidence, to report the development and construct validation of a tool for measuring the effectiveness of principal and school-wide leadership. While other measures of instructional leadership are available (Goldring *et al.*, 2009a; Hallinger and Murphy, 1985), the particular contribution of this study is to identify, through a series of Rasch analyses, the progression of difficulty in performing effectively on the practices described by the items in each of the eight scales. In the remainder of this section we discuss limitations of this study and future research, before turning to possible implications of our study for principal, school and leadership development.

In this study we have established the construct and not the predictive validity of the ELP tool. Establishing the latter requires research that examines relationships through either correlational or intervention studies, between leadership effectiveness on the ELP and theoretically relevant teacher, school or student outcomes measures. The stability of the ELP over time makes it suitable for the evaluation of leadership development interventions, as any resulting changes are unlikely to be caused by measurement unreliability. Given the current emphasis on building instructional leadership capability in principals and school leadership teams, the use of this tool in the evaluation of leadership development seems warranted. Studies are also needed to establish the concurrent validity of the tool by, for example, analyzing its relationship to other available measures of instructional leadership.

The ELP provides principals with rich data, including a scale by scale comparison of their own and their teachers' perceptions of principal and school-wide leadership effectiveness. The perceptual rather than entirely objective measurement approach suggests a need for attention to commonly reported limitations of perceptual data. The social psychology literature would suggest that data based on own and other's perceptions of effectiveness are likely to be influenced by positive illusions (Taylor and Brown, 1988) whereby individuals (in this case principals) have highly skewed positive views of themselves, and by self-serving bias (Miller *et al.*, 1975) whereby teachers rating principals might attribute their own failures to the principal while attributing their successes to themselves. Findings from prior research (Sinnema *et al.*, 2015) investigating the discrepancy between principal and teacher ratings on one of the ELP scales (principal effectiveness) suggest that the impact of these psychological processes on the principal effectiveness scale of the ELP tool at least, are not as significant as they may be in other contexts. That study of discrepancy revealed that while both principals and teachers rate the principal highly (on principal effectiveness) teachers tended to rate their principal higher than the principals rated themselves. Illusion therefore, was

possibly at play to some extent for both respondent groups, but no more so for those rating themselves (principals) than for those rating others (teachers). Findings from the same study about variables associated with greater magnitudes of discrepancy do signal that any limitations of perceptual measurement are particularly pertinent in the use of this tool to rate principals who are younger, have had less time in the role of principal at the school, and who are leading schools of lower socio-economic status.

Notwithstanding the presence and inevitability of at least some discrepancies, and the limitation of this study in not examining the impact of these psychological constructs for all scales, we consider there to be much value for school leaders in engaging with data generated from this tool. The Rasch variable maps provide principals with individual profiles of their perceived effectiveness on each scale. The progressive difficulty of the items means that the profile not only identifies the more and less effective leadership practices, but also suggests the appropriate next steps in that principal's leadership development. It enables principals to see their leadership learning as a developmental progression and to set professional goals that focus on practices that are slightly more difficult than those they have already mastered. The profile of school-wide leadership effectiveness provides a diagnostic picture of instructional leadership across the school and, can be used to identify priority areas for improvement and next steps for school development.

When aggregated across a system, the ELP profiles enable precise mapping of the stretch between current leadership capability and that required to achieve system goals. Diagnosing leadership effectiveness from a system perspective means that resources can be allocated to target those practices that are most lacking in the sector. This is important, since it optimizes the chance of improvement, and reduces the unnecessary provision of interventions focussed on practices where there is already widespread capability.

The Rasch analysis reported here gives valuable insights into the progressive difficulty of many of the practices involved in instructional leadership. The progression from more routine to more rigorous outcome-focussed leadership has important implications for system-wide approaches to leadership development. The relative difficulty of practices that deal with specific problems (including conflict resolution and problem solving) signals the importance of developing school leaders' interpersonal capability. It also signals the need to ensure that evaluations of interventions to improve interpersonal capability assess impact on actual problem resolution as well as on leaders' interpersonal skills. Similarly, our analyses indicate the need for greater attention to how school leaders might more effectively address diversity in their leadership efforts. While items in the ELP refer specifically to Māori students, and that specificity may be seen to limit the generalizability of this tool internationally, our results suggest the need for a sustained emphasis on understanding and overcoming the challenges involved in serving diverse student populations, and indigenous students in particular.

The relative difficulty of the items about engaging relevant expertise has important implications and needs further investigation. Does it signal the paucity of expertise, difficulty in accessing it or low capability in determining the type of expertise that is needed? The need for school leaders to draw on expertise to support improvement in context-specific problems of practice seems particularly pressing given indications that improvement-focussed leadership practices are also at the harder end of the spectrum.

This study suggests that principals, like the students and teachers they serve, may be understood to develop skills and capabilities along (or up, as the Rasch variable maps suggest) continua of practices. The development involved in these practices may

be measured and characterized as a set of sequences that offer differing entry levels for leaders at varying stages of development on any one of the dimensions. The prospect of personalized and needs-based leadership development become more possible if the tool is used in a diagnostic and developmental way. With more targeted provision, improvements in leadership practice are more likely to influence both the quality of teaching and ultimately outcomes for students. To establish the extent to which such benefits are realized, it would be desirable for future research to link measures of leadership practice (and measures of intervention impact on leadership practice) to measures of instructional quality and student learning.

Acknowledgements

The ELP was developed by the New Zealand Council of Educational Research under contract to the national Ministry of Education. The third author was involved at the stage of item writing. The authors wish to thank Andrew Porter (University of Pennsylvania), Joseph Murphy (Vanderbilt University), Ellen Goldring (Vanderbilt University), and Stephen N. Elliott (Arizona State University), the authors of the Vanderbilt Assessment of Leadership in Education, for use of portions of the Vanderbilt Assessment of Leadership in Education (VAL-ED) structure and other items from the VAL-ED.

The authors also wish to express their gratitude to the New Zealand Council for Educational Research for their ongoing assistance following the administration of the ELP, and for the valuable feedback from Graeme Cosslett and Cathy Wylie on draft material.

Note

1. From this point the terms "easiest" or "easier" will be used to refer to "easiest/easier to rate as outstandingly effective" and "hardest" or "harder" will be used to refer to "hardest /harder to rate as outstandingly effective".

References

Andrich, D. (1978), "A rating formulation for ordered response categories", *Psychometrika*, Vol. 43 No. 4, pp. 561-573.

Antoniou, P. (2013), "Development of research on school leadership through evidence-based and theory driven approaches: a review of school leadership effects revisited", *School Effectiveness and School Improvement*, Vol. 24 No. 1, pp. 122-128.

Borman, G.D., Hewes, G.M., Overman, L.T. and Brown, S. (2003), "Comprehensive school reform and achievement: a meta-analysis", *Review of Educational Research*, Vol. 73 No. 2, pp. 125-230.

Bryk, A.S. and Schneider, B.L. (2002), *Trust in Schools: A Core Resource for Improvement*, Russell Sage Foundation Publications, New York, NY.

Datnow, A. and Park, V. (2014), *Data-Driven Leadership*, Jossey Bass, San Francisco, CA.

De Ayala, R.J. (2009), *The Theory and Practice of Item Response Theory*, Guilford Press, New York, NY.

Education Review Office (2013), "Increasing educational achievement in secondary schools", New Zealand Government, Wellington, NZ.

Embretson, S.E. and Reise, S.P. (2000), *Item Response Theory for Psychologists*, Lawrence Erlbaum Associates, Mahwah, NJ.

Epstein, J.L. and Sheldon, S. (2006), "Moving forward: Ideas for research on school, family, and community partnerships", in Conrad, C.F. and Serlin, R. (Eds), *SAGE Handbook for*

Research in Education: Engaging Ideas and Enriching Inquiry, Sage Publications, Thousand Oaks, CA, pp. 117-137.

Friedkin, N.E. and Slater, M.R. (1994), "School leadership and performance: a social network approach", *Sociology of Education*, Vol. 67 No. 2, pp. 139-157.

Goddard, R.D., Salloum, S.J. and Berebitsky, D. (2009), "Trust as a mediator of the relationships between poverty, racial composition, and academic achievement: evidence from Michigan's public elementary schools", *Educational Administration Quarterly*, Vol. 45 No. 2, pp. 292-311.

Goldring, E.B., Huff, J., Spillane, J.P. and Barnes, C. (2009a), "Measuring the learning-centered leadership expertise of school principals", *Leadership and Policy in Schools*, Vol. 8 No. 2, pp. 197-228.

Goldring, E.B., Porter, A.C., Murphy, J., Elliott, S.N. and Cravens, X. (2009b), "Assessing learning-centered leadership: connections to research, professional standards and current practices", *Leadership and Policy in Schools*, Vol. 8 No. 1, pp. 1-36.

Grissom, J.A. (2011), "Can good principals keep teachers in disadvantaged schools? Linking principal effectiveness to teacher satisfaction and turnover in hard-to-staff environments", *Teachers College Record*, Vol. 113 No. 11, pp. 2552-2585.

Grissom, J.A. and Loeb, S. (2011), "Triangulating principal effectiveness", *American Educational Research Journal*, Vol. 48 No. 5, pp. 1091-1123.

Grissom, J.A., Loeb, S. and Master, B. (2013), "Effective instructional time use for school leaders: longitudinal evidence from observations of principals", *Educational Researcher*, Vol. 42 No. 8, pp. 433-444.

Hallinger, P. and Heck, R.H. (2002), "What do you call people with visions? The role of vision, mission, and goals in school leadership and improvement", in Leithwood, K. and Hallinger, P. (Eds), *Second International Handbook of Educational Leadership and Administration*, Kluwer, Dordrecht, pp. 9-40.

Hallinger, P. and Murphy, J. (1985), "Assessing the instructional management behavior of principals", *Elementary School Journal*, Vol. 86 No. 2, pp. 217-247.

Hattie, J. (1984), "An empirical study of the various indices for determining unidimensionality", *Multivariate Behavioral Research*, Vol. 19 No. 1, pp. 49-78.

Heck, R.H. (2000), "Examining the impact of school quality on school outcomes and improvement: a value-added approach", *Educational Administration Quarterly*, Vol. 36 No. 4, pp. 513-552.

Heck, R.H., Larsen, T.J. and Marcoulides, G.A. (1990), "Instructional leadership and school achievement: validation of a causal model", *Educational Administration Quarterly*, Vol. 26 No. 2, pp. 94-125.

Heck, R.H., Marcoulides, G.A. and Lang, P. (1991), "Principal instructional leadership and school achievement: the application of discriminant techniques", *School Effectiveness and School Improvement*, Vol. 2 No. 2, pp. 115-135.

Heinrich, M. (2012), "Bridging accountability obligations, professional values and (perceived) student needs with integrity", *Journal of Educational Administration*, Vol. 50 No. 5, pp. 695-726.

Kaiser, H.F. (1970), "A second generation little jiffy", *Psychometrika*, Vol. 35, pp. 401-415.

Kaiser, H.F. and Rice, J. (1974), "Little jiffy, mark iv", *Educational and Psychological Measurement*, Vol. 34 No. 1, pp. 111-117.

Karabenick, S., Woolley, M., Friedel, J., Ammon, B., Blazevski, J., Bonney, C.R., Groot, E.D., Gilbert, M., Musu, L., Kempler, T. and Kelly, K. (2007), "Cognitive processing of self-report items in educational research: do they think what we mean?", *Educational Psychologist*, Vol. 42 No. 3, pp. 139-151.

Leithwood, K., Harris, A. and Hopkins, D. (2008), "Seven strong claims about successful school leadership", *School Leadership & Management*, Vol. 28 No. 1, pp. 27-42.

Linacre, J.M. (2012), "WINSTEPS (Version 3.75.1)", computer program), Winsteps.com, Beaverton, OR.

Ludlow, L.H. (1983), "The analysis of Rasch model residuals", unpublished doctoral dissertation, University of Chicago, Chicago, IL.

Ludlow, L.H. (1985), "A strategy for the graphical representation of Rasch model residuals", *Educational and Psychological Measurement*, Vol. 45 No. 4, pp. 851-859.

Ludlow, L.H. (1986), "Graphical analysis of item response theory residuals", *Applied Psychological Measurement*, Vol. 10 No. 3, pp. 217-229.

Ludlow, L.H. and Haley, S.M. (1995), "Rasch model logits: interpretation, use, and transformation", *Educational and Psychological Measurement*, Vol. 55 No. 6, pp. 967-975.

Ludlow, L.H., Matz-Costa, C., Johnson, C., Brown, M., Besen, E. and James, J.B. (2014), "Measuring engagement in later life activities: Rasch-based scenario scales for work, caregiving, informal helping, and volunteering", *Measurement and Evaluation in Counseling and Development*, Vol. 47 No. 2, pp. 127-149.

May, S., Cowles, S. and Lamy, M. (2012), "PISA 2012: New Zealand summary report", Comparative Education Research Unit, Wellington, NZ.

Miles, K.H. and Frank, S. (2008), *The Strategic School: Making the Most of People Time and Money*, Corwin Press, Thousand Oaks, CA.

Miller, D.T., Dale, T. and Ross, M. (1975), "Self-serving biases in attribution of causality: fact or fiction?", *Psychological Bulletin*, Vol. 82 No. 2, pp. 213-225.

Ministry of Education (2008), "Kiwi leadership for principals. Principals as educational leaders", Ministry of Education, Wellington, NZ.

O'Connor, B.P. (2000), "SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test", *Behavior Research Methods Instruments and Computers*, Vol. 32 No. 3, pp. 396-402.

Orr, M.T. and Orphanos, S. (2011), "How graduate-level preparation influences the effectiveness of school leaders: a comparison of the outcomes of exemplary and conventional leadership preparation programs for principals", *Educational Administration Quarterly*, Vol. 47 No. 1, pp. 18-70.

Presser, S. and Blair, J. (1994), "Survey pretesting – do different methods produce different results?", *Sociological Methodology*, Vol. 24, pp. 73-104.

Rasch, G. (1960), *Probabilistic Models for Some Intelligence and Attainment Tests*, The University of Chicago Press, Chicago, MI.

Reckase, M.D. (1979), "Unifactor latent trait models applied to multifactor tests: results and implications", *Journal of Educational Statistics*, Vol. 4 No. 3, pp. 207-230.

Reise, S.P. and Revicki, D.A. (Eds) (2015), *Handbook of Item Response Theory Modeling*, Routledge, New York, NY.

Robinson, V.M.J. (2007), "School leadership and student outcomes: identifying what works and why", Australian Council for Educational Leaders, Winmalee.

Robinson, V.M.J. (2011), *Student-Centered Leadership*, Jossey Bass, San Francisco, CA.

Robinson, V.M.J., Hohepa, M. and Lloyd, C. (2009), "School leadership and student outcomes: identifying what works and why. Best evidence synthesis iteration", New Zealand Ministry of Education, Wellington, NZ.

Robinson, V.M.J., Lloyd, C. and Rowe, K.J. (2008), "The impact of leadership on student outcomes: an analysis of the differential effects of leadership type", *Educational Administration Quarterly*, Vol. 44 No. 5, pp. 635-674.

Rollison, J.M., Ludlow, L.H. and Wallingford, T. (2012), "Assessing content knowledge and changes in confidene and anxiety related to economic literacy in a profesional development program for history teachers", *Journal of Educational Research and Practice*, Vol. 2 No. 1, pp. 15-30.

Scherbaum, C.A., Finlinson, S., Barden, K. and Tamanini, K. (2006), "Applications of item response theory to measurement issues in leadership research", *The Leadership Quarterly*, Vol. 17 No. 4, pp. 366-386.

Seijts, G.H. and Latham, G.P. (2012), "Knowing when to set learning versus performance goals", *Organizational Dynamics*, Vol. 41 No. 1, pp. 1-6.

Slocum, S.L. (2005), "Assessing unidimensionality of psychological scales: using individual and interpretive criteria from factor analysis", unpublished doctoral thesis, The University of British Columbia, Vancouver.

Sinnema, C. and Ludlow, L.H. (2013), "A Rasch approach to the measurement of responsive curriculum practice in the context of curricula reform", *International Journal of Educational and Psychological Assessment*, Vol. 12 No. 2, pp. 33-40.

Sinnema, C. and Robinson, V.M.J. (2007), "The leadership of teaching and learning: implications for teacher evaluation", *Leadership and Policy in Schools*, Vol. 6 No. 4, pp. 319-343.

Sinnema, C. and Robinson, V.M.J. (2012), "Goal setting in principal evaluation: goal quality and predictors of achievement", *Leadership and Policy in Schools*, Vol. 11 No. 2, pp. 135-167.

Sinnema, C., Robinson, V.M.J., Ludlow, L.H. and Pope, D. (2015), "How effective is the principal? Discrepancy between teachers' and principals' perceptions of principal effectiveness", *Educational Assessment, Evaluation and Accountability*, Vol. 27 No. 3, pp. 275-301.

Smith, R.M. (1991), "The distributional properties of Rasch item fit statistics", *Educational and Psychological Measurement*, Vol. 51 No. 3, pp. 541-565.

Smith, R.M., Schumacker, R.E. and Bush, M.J. (1998), "Using item mean squares to evaluate fit to the Rasch model", *Journal of Outcome Measurement*, Vol. 2 No. 1, pp. 66-78.

Stout, W. (1987), "A nonparametric approach for assessing latent trait unidimensionality", *Psychometrika*, Vol. 52 No. 4, pp. 589-617.

Taylor, S.E. and Brown, J. (1988), "Illusion and well-being: a social psychological perspective on mental health", *Psychological Bulletin*, Vol. 103 No. 2, pp. 193-210.

Timperley, H. and Alton-Lee, A. (2008), "Reframing teacher professional learning: an alternative policy approach to strengthening valued outcomes for diverse learners", *Review of Research in Education*, Vol. 32 No. 1, pp. 328-369.

Timperley, H., Wilson, A., Barrar, H. and Fung, I. (2007), "Teacher professional learning and development: best evidence synthesis iteration", Ministry of Education, Wellington.

Vescio, V., Ross, D. and Adams, A. (2008), "A review of research on the impact of professional learning communities on teaching practice and student learning", *Teaching and Teacher Education*, Vol. 24 No. 1, pp. 80-91.

Wang, M.-T. and Holcombe, R. (2010), "Adolescents' perceptions of school environment, engagement and academic achievement in middle school", *American Educational Research Journal*, Vol. 47 No. 3, pp. 633-662.

Wright, B.D. and Masters, G.N. (1982), *Rating Scale Analysis*, MESA Press, Chicago, IL.

Wright, B.D. and Stone, M.H. (1979), *Best Test Design*, MESA Press, Chicago, IL.

Wright, B.D. and Panchapakesan, N. (1969), "A procedure for sample-free item analysis", *Educational and Psychological Measurement*, Vol. 29 No. 1, pp. 23-48.

Further reading

Latham, G.P. and Locke, E.A. (2006), "Enhancing the benefits and overcoming the pitfalls of goal setting", *Organizational Dynamics*, Vol. 35 No. 4, pp. 332-340.

Locke, E.A. and Latham, G.P. (1990), *A Theory of Goal Setting and Task Performance*, Prentice Hall, Englewood Cliffs, NJ.

Marzano, R.J., Waters, T. and Mcnulty, B. (2005), *School Leadership That Works: From Research to Results*, ASCD and McREL, Aurora, CO.

About the authors

Claire Sinnema is a Senior Lecturer in the Faculty of Education at the University of Auckland, New Zealand. Her research focusses on the improvement of teaching and learning across four main strands – curriculum implementation, teacher professional learning, pedagogy and school leadership. It is concerned with understanding how teachers and principals can improve their practice in ways that enhance student learning. It is also concerned with the role of policy in such improvement. Claire Sinnema is the corresponding author and can be contacted at: c.sinnema@auckland.ac.nz

Larry Ludlow is a Professor and the Chair of the Department of Educational Research, Measurement and Evaluation in the Lynch School of Education at Boston College. He teaches courses in research methods, statistics and psychometrics. His research interests include development of: longitudinal survey and data analysis systems for tracking teacher candidates, longitudinal models for understanding and predicting faculty teaching evaluations, and longitudinal teacher retention and attrition prediction models.

Viviane Robinson is a Distinguished Professor in the Faculty of Education at the University of Auckland and the Academic Director of its Centre for Educational Leadership. As an organizational psychologist working in the field of education she pursues a research program which is concerned with organizational learning and leadership, school effectiveness and improvement, and research methodology.