Educ Asse Eval Acc (2017) 29:211–224
DOI 10.1007/s11092-016-9254-9

A test-retest analysis of the Vanderbilt Assessment for Leadership in Education in the USA

Elizabeth Covay Minor[1,2,3] · **Andrew C. Porter**[2] · **Joseph Murphy**[4] · **Ellen Goldring**[4] · **Stephen N. Elliott**[5]

Received: 14 September 2015 / Accepted: 11 November 2016 / Published online: 23 November 2016

Abstract The Vanderbilt Assessment for Leadership in Education (VAL-ED) is a 360-degree learning-centered behaviors principal evaluation tool that includes ratings from the principal, supervisors, and teachers. The current study assesses the test-retest reliability of the VAL-ED for a sample of seven school districts as part of multiple validity and reliability assessments based on various samples of real users of the VAL-ED. We administered the VAL-ED twice and examined the correlations and mean differences between time 1 and time 2. We find that the principal and teacher ratings from time 1 and time 2 have large, positive, and significant correlations. Additionally,

✉ Elizabeth Covay Minor
eminor1@nl.edu

Andrew C. Porter
andyp@gse.upenn.edu

Joseph Murphy
Joseph.f.murphy@vanderbilt.edu

Ellen Goldring
ellen.goldring@vanderbilt.edu

Stephen N. Elliott
steve_elliott@asu.edu

[1] National Louis University, 1000 Capitol Drive, Wheeling, IL 60090, USA

[2] University of Pennsylvania, 3700 Walnut Street, Philadelphia, PA 19104, USA

[3] Michigan State University, East Lansing, MI, USA

[4] Vanderbilt University, Peabody College, Nashville, TN 37203, USA

[5] Sanford School of Social and Family Dynamics, Arizona State University, 308K Payne Hall East, Tempe, AZ 85287-7805, USA

for both time points, principals are rated as being at least satisfactorily effective. Principals rate themselves slightly higher at time 2, while teachers rate principals slightly higher at time 1.

Keywords Learning-centered leadership · Test-retest reliability · Principal evaluation

1 A test-retest analysis of the Vanderbilt Assessment for Leadership in Education in the USA

Researchers have continually found that principals are an important component of school effectiveness (Leithwood and Montgomery 1982; Murphy et al. 1983, 1985; Bryk et al. 2010), especially for good teaching and strong student learning (Leithwood and Sun 2012). A common characteristic of effective principals is the use of learning-centered leadership behaviors (Leithwood et al. 2004; Leithwood et al. 2006; Marks and Printy 2003; May and Supovitz 2011; Preston et al. 2012; Robinson et al. 2008; Supovitz et al. 2009), which refers to the school leaders staying focused on the manifest function of schooling—learning. This includes making sure that various actors within the schooling community work together to ensure student learning (Murphy et al. 2007).

Given the accountability and effective schools movement, it is of critical importance to have a high quality measure of learning-centered leadership behaviors. For example, the Race to the Top (RTTT) Program asks states to institute reform in order to improve student outcomes. RTTT focuses on four areas: (1) college and career ready standards, (2) systems to monitor student achievement growth, (3) ensuring schools have and retain effective teachers and principals, and (4) reform the lowest achieving schools. The Vanderbilt Assessment for Leadership in Education (VAL-ED) is a tool that can help address the need for a sound measure of learning-centered leadership behaviors overcoming the weaknesses of many principal evaluation systems (Goldring et al. 2009b).

The VAL-ED was developed to address the weakness of other principal evaluation systems. School systems can use principal evaluations as a way to influence principal learning-centered leadership behaviors (Murphy 1989, 1992; Murphy et al. 2012) and in turn the effectiveness of their schools. However, there are weaknesses in many of the current principal evaluation systems (for a more thorough examination of these weaknesses, see Goldring et al. 2009b). Namely, most systems do not adhere to the professional standards (AERA, APA, NCME 1999) for testing, and they are not well grounded in research on effective leadership and school improvement. The VAL-ED is grounded in leadership research, aligned to the Interstate School Leaders Licensure Consortium (ISLLC) standards, and increasingly documented as psychometrically sound. What follows is documentation of the instrument's test-retest reliability for a sample of seven school districts.

To adhere to testing standards, the VAL-ED has been subjected to rigorous investigation of its psychometric properties (e.g., reliability and validity on a national representative sample of schools in the USA). During the development phase of the VAL-ED, our research team conducted a sorting study, cognitive labs, an item bias

study, and two pilot studies (Porter et al. 2008). The instrument was then tested on a larger scale through a US nationally representative field trial in 2008 (see Porter et al. 2010 for more information).

Presently, the research team is undertaking additional validity and reliability studies for multiple samples of real users of the VAL-ED. In other words, rather than recruiting schools for the expressed purpose of research and asking them to take the VAL-ED, the real user studies recruited schools that are already using the VAL-ED for their own purposes. These studies include a known group analysis (Covay et al. 2014), a convergent-divergent study (Goldring et al. 2015), a study on how schools use results, and a study of the utility of VAL-ED in predicting value added of principal leadership on student achievement.

The results so far do show evidence that the VAL-ED is a valid and reliable principal evaluation tool for those samples of real users examined as well as in research sites for the US nationally representative sample. For example, we conducted a known group study on real users in six school districts, which requires superintendents to list the principals in the top 20% of their district and in the bottom 20% of their district. We found that from those six districts, those principals in the top 20% have significantly higher VAL-ED scores compared to those principals in the bottom 20% of the district. The VAL-ED was able to distinguish between these two groups of principals (Covay et al. 2014). The convergent-divergent study (Goldring et al. 2015), which included eight school districts, found that in those districts, principal self-ratings on the VAL-ED are convergent with another leadership scale—the Principal Instructional Management Rating Scale (Hallinger and Murphy 1985; Hallinger 2011; Hallinger et al. 2013)—and divergent on a measure of emotional intelligence—the Trait Emotional Intelligence Questionnaire (Petrides et al. 2007).

The current study makes a unique contribution through investigation of the test-retest reliability of the instrument on a sample of real users from seven school districts. The purpose of the test-retest study is to assess the extent to which results from the VAL-ED are consistent over time.

2 VAL-ED conceptual framework

The conceptual model of the VAL-ED begins with principal's knowledge and skills, personal characteristics, and values and beliefs, which contribute to the principal's leadership behaviors. These leadership behaviors, which are described in more detail below, lead to better instruction which in turn lead to increase in student success. The VAL-ED focuses on the leadership behaviors from the conceptual model (Porter et al. 2010). More specifically, the VAL-ED looks at the intersection of two dimensions: *core components* and *key processes*. The core components—or what the school leader must do to increase the academic and social learning of students—are created through key processes—how they go about creating those components needed to increase student academic and social learning. The identification of the core components and key processes came through an extensive literature review on learning-centered leadership (Goldring et al. 2009a) and aligned with the dimensions of learning-centered leadership that past research has identified (Murphy et al. 2007). Not only do the core components

and key processes align with previous research on learning-centered leadership behaviors but also the ISLLC standards (Goldring et al. 2009a).

There are six core components, which include high standards for student performance, rigorous curriculum (content), quality instruction (pedagogy), culture of learning and professional behavior, connections to external communities, and systemic performance accountability. *High standards* for student performance include having rigorous goals for both academic and social learning and related to vision and mission. The core component of a *rigorous curriculum* means that principals ensure that students receive challenging instructional content, while *quality instruction* refers to pedagogy that maximizes student learning and depth of understanding. Principals should also help to create a *culture of learning and professional behavior*, which means that the central focus of the school is student learning. This includes learning communities and a positive school culture related to student learning. Additionally, effective principals maintain *connections to external community* actors such as parents and community partners to improve student success. The final core component is systematic *performance accountability* which includes individual and collective responsibility for student learning among all school actors. This includes both internal and external accountability (Goldring et al. 2009a; Porter et al. 2010).

There are six key processes: planning, implementing, supporting, advocating, communicating, and monitoring. *Planning* involves clearly articulating a shared mission/goals and establishing the necessary steps and procedures to achieve those goals. *Implementing* involves putting plans into action. Effective instructional leaders *support* the achievement of these goals by creating an environment conducive for student learning. *Advocating* involves making sure that the needs of all students are met. *Communicating* is the process of establishing and maintaining avenues for exchange among school actors. Finally, *monitoring* means that the principal collects information to evaluate that progress is being made in achieving the core components (Goldring et al. 2009a; Porter et al. 2010).

3 VAL-ED instrument

The VAL-ED includes ratings from the principal him/herself, the principal's supervisor, and the teachers within the school to form a 360-degree assessment. It is considered a 360-degree assessment because the principal is being evaluated from multiple directions. Each respondent rates the principal's effectiveness on two behaviors from each of the core component- key processes combination (i.e., 36 domains). Most respondents completed the VAL-ED online; however, paper and pencil versions are possible (for more information on the validity and reliability from the US national field trial, see Porter et al. 2010).

Before evaluating the principal's effectiveness, the respondent is first asked to select the sources of their evidence that they use to make the evaluation. Respondents select all that apply from "reports from others," "personal observations," "school documents," "school projects or activities," "other sources," or "no evidence." Next, the respondent rates the principal's effectiveness for each behavior on a scale from 1 to 5: 1 = ineffective, 2 = minimally effective, 3 = satisfactorily effective, 4 = highly effective, and 5 = outstandingly effective. There is an option for teachers and

supervisors to select "don't know" for an effectiveness rating; however, this is not an option for principals given that they should know whether or not they completed a behavior. There are a few instances when the source of evidence or lack of evidence limits respondents' effectiveness rating options. If teachers or supervisors have "no evidence" of a behavior, then they can select "ineffective" or "don't know" as the effectiveness rating. However, if a principal selects "no evidence," then they must select "ineffective" as this indicates that the behavior was not completed.

For each principal, a total effectiveness score is calculated along with a subscore for each of the core components and key processes. In calculating the scores, the principal's response, supervisor's response, and average teachers' responses are combined with each group weighted equally. In addition to reporting the 360 degree, total effectiveness score and subscores are reported by principal, supervisor, and/or teachers. After completion of the VAL-ED, the principal receives a report that includes his/her scores on the 5-point scale, percentile ranks based on national norms and performance levels for the USA (Polikoff et al. 2009; Porter et al. 2010).

4 Reliability

The VAL-ED underwent a US national field trial in 2008 to assess the reliability and validity of the instrument for a sample of 300 nationally representative schools (Porter et al. 2010). That work examined reliability and validity (including bias), for the nationally representative sample, in regard to total and scale scores, equivalence of parallel forms, and established norms and performance standards. The national field trial showed evidence of high internal consistency reliability, which is a prerequisite for reliability, and constructed validity for the nationally representative sample of the USA (Cravens et al. 2013; Porter et al. 2010; Polikoff et al. 2009); however, the national field trial results were based on research sites recruited for the purpose of studying the VAL-ED. In subsequent work, the psychometric properties of the VAL-ED for real users (i.e., districts that have purchased the VAL-ED for their own principal evaluations) are being assessed (Goldring et al. 2015; Covay et al. 2014). It is possible that schools or districts that seek out and purchase the VAL-ED may be different than those that were recruited as part of the Wallace funded national field trial. The real users of the VAL-ED would have motivations for including a principal evaluation tool in their district, whereas the research sites may not have had these motivations in mind. This study takes an important step toward filling that knowledge gap and developing an argument of validity and reliability for various samples of real users for the VAL-ED.

This particular study focuses on the test-retest reliability of the VAL-ED for a sample of seven school districts. More specifically, we ask: To what extent are VAL-ED results consistent at two points in time and is this consistent by the role of the rater (teacher, self, supervisor)?

The test-retest method is a classic test of reliability, involving the correlation of a measure across two time points (Carmine and Zeller 1979). It compliments internal

consistency reliability by reporting on stability over time for a given population of real users.

5 Data and methods

We recruited schools using the VAL-ED to participate in the test-retest study. The test-retest study design required that principals, supervisors, and teachers take the VAL-ED twice within a four-week period once as a part of regular use and a second time for our research purposes. Depending on the district, the first administration took place during the months of September through January with the second administration occurring mostly from November through February of the 2011–2012 school year. One district had their second administration in May and one district completed both administrations in the 2010–2011 school year. Included in the test-retest study are a total of 71 schools (both elementary and secondary schools) across seven school districts. While most of the districts came from the Midwest, most of the schools came from the South. The districts from the Midwest range from 4 to 17 schools, but the district from the South had close to 170 schools. Not all of the schools were included in the analyses, since not all of the schools had completed two administrations of the VAL-ED.[1] Across the respondent groups, 35 principals took the VAL-ED twice, 7 supervisors, and teachers from 71 schools. We limit our discussion of results to the principals and teachers, since so few supervisors took the VAL-ED twice.

For each school, a total effectiveness score along with individual core components and key processes subscale ratings was calculated for each respondent group. The total effectiveness score is an aggregate value of the six core components (or the six key processes) subscales.

To examine the test-retest reliability of the VAL-ED for this particular sample, we performed two sets of analyses. The first analysis was a correlation between time 1 and time 2 for the total effectiveness score, the six core components, and the six key processes. The second analysis was a mean difference analysis comparing the differences between time 1 and time 2 via paired sample *t* tests. We would expect significant correlations between time 1 and time 2, but not significant mean differences between the two administrations. The lag between time 1 administration and time 2 administration varied from 2 and a half weeks to 29 weeks. We ran the analyses with all schools as well as only the schools that had a 2–8-week lag to reduce variation due to long lags in the testing. We compared the results of the limited sample with the 2–8-week lag to the full sample, which we use for our final results. The significance levels and the magnitudes of the correlations differed slightly between the two analyses; however, the patterns are consistent. For the most part, the results from the shorter lag period had smaller correlations and lower means. For example, the correlation for total effectiveness for principal self-ratings at time 1 and time 2 in districts with a lag of 2–8 weeks is 0.6051 ($N = 19$) compared to our analytic sample with 0.6392 ($N = 35$). The correlation for those with a lag greater than 8 weeks is 0.4830 ($N = 15$). The correlation for teacher

[1] The total principal effectiveness ratings for time 1 for those principals who completed the time 2 rating (3.66) were higher compared to those who did not complete the time 2 assessment (3.52), but the differences are not significant.

reports is also slightly lower for the districts with a lag of 2–8 weeks (r = 0.9075; N = 37) compared to our analytic sample (r = 0.9092; N = 71) but slightly higher than those in districts with greater than an 8-week lag (r = 0.9070; N = 34). The mean principal total effectiveness self-rating at time 1 for those districts with a lag of 2–8 weeks is 3.42 and at time 2 is 3.66 compared to our analytic sample of principal self-rating at time 1 of 3.66 and time 2 of 3.89. For those in a district with a lag greater than 8 weeks, the principal self-rating at time 1 is 3.77 and at time 2 is 4.16. There are also slight differences for teacher ratings of principals for the 2–8-week lag (time 1 = 3.60; time 2 = 3.48), our analytic sample (time 1 = 3.67; time 2 = 3.58), and those with a lag greater than 8 weeks (time 1 = 3.74; time 2 = 3.68).

6 Results

6.1 Correlations

The correlation between time 1 *total effectiveness* and time 2 *total effectiveness* for principal self-ratings is positive, large[2] and significant (r = 0.6392; r^2 = 0.4086). The correlations between the two administrations for the core components and key processes are also positive, large, and significant for principal self-ratings (see Table 1). The magnitude of the correlations for the core components ranges from a low of 0.5510 (r^2 = 0.3036) for the core component of *external community* to a high of 0.7092 (r^2 = 0.5030) for *high standards*. The range of magnitudes for the correlations for the principal's self-ratings of key processes is slightly narrower than those for the core components. The key process with the lowest correlation between time 1 and time 2 is *advocating* with a correlation of 0.5559 (r^2 = 0.3090), and the key process with the highest correlation is *planning* with a correlation of 0.6976 (r^2 = 0.4866). All of these are considered large correlations (Cohen 1988; Fan 2001).

As in the case of principal self-reports, the correlations for teacher reports at time 1 and time 2 are positive, slightly larger, and significant. The correlation between time 1 and 2 for *total effectiveness* is 0.9092 (r^2 = 0.8266). The largest correlation is for the key process of *planning* at 0.9212 (r^2 = 0.8486), and the key process with the lowest correlation between time 1 and time 2 is *communicating* at 0.8738 (r^2 = 0.7635). The core component with the largest correlation is *high standards* at 0.9125 (r^2 = 0.8327), and the smallest is for the core component of *external community* at 0.8731 (r^2 = 0.7623). The higher test-retest reliability for teacher data than for the principal data may be a function of the teacher data being based on averages. Averaging a larger number of ratings (i.e., teachers) provides a more reliable estimate compared to relying on the rating of one person (i.e., the principal), since the average can reduce noise in the estimate (National Instrument 2006). The response rate for teacher per school average was about 69% with a range from 6 to 100%. Time 1's average response rate was 77% and time 2's average response rate was 60%.

[2] In the discussion of the magnitude of correlations, we use the rules of thumb from Cohen (1988) [see Fan (2001) for an application] where a r^2 of 0.02 is a small correlation, r^2 of 0.12 is a medium correlation, and r^2 of 0.25 is a large correlation.

Table 1 Correlations between time 1 and time 2 VAL-ED scores

	Principal (N = 35)	Teachers (N = 71)
	Total effectiveness and core components	
Total	0.6392***	0.9092***
High standards	0.7092***	0.9125***
Rigorous curriculum	0.6365***	0.8877***
Quality instruction	0.5573***	0.8731***
Culture of learning	0.6048***	0.9068***
External community	0.5510***	0.8693***
Performance accountability	0.6254***	0.9038***
	Key processes	
Planning	0.6976***	0.9212***
Implementing	0.6573***	0.9009***
Supporting	0.5736***	0.8925***
Advocating	0.5559***	0.9221***
Communicating	0.6242***	0.8738***
Monitoring	0.6191***	0.8901***

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

We redid these analyses with a sample that had at least 50% teacher response rate. The correlations were slightly higher in magnitude for the sample restricted by the response rate. The magnitude for the mean differences changed little. When comparing the mean values for the principal reported VAL-ED scores, the restricted sample (N = 25) yielded slightly lower mean (but at most 0.09 lower). Additionally, the teacher (N = 46) mean values for the VAL-ED scores changed little. Although most means were slightly higher (at most 0.06) for the restricted sample, some values stayed the same and some were slightly lower in the restricted same (at most 0.02). The significance levels for some comparisons did change, mostly when looking at the teacher reports but some for the principal reports. For the most part, we see that the restricted sample has fewer significant differences for the mean comparisons, which is likely due to reduced power from a smaller sample size. However, the overall patterns remained. Teachers score principals lower at time 2 and principals score themselves higher at time 2. This could be a testing effect like that described by Campbell and Stanley (1963).

Overall, we see that the total effectiveness scores, core components, and key processes at time 1 are significantly correlated with the time 2 administration.

6.2 Mean differences

Another way to examine the relationship between time 1 administration of the VAL-ED and time 2 administration is to examine whether or not there are significant mean differences between the two administrations. Unexpectedly, the mean differences between time 1 and time 2 administrations for principals' self-reports are significantly different from each other with the exception of the core component of *quality instruction* and the key process of *supporting* (see Table 2). Consistently, principals rate

Table 2 Principal level means

	Time 1			Time 2		
	Mean	SD	95% Confidence intervals	Mean	SD	95% Confidence intervals
Total score	3.66**	0.483	3.49–3.82	3.89	0.594	3.69–4.09
Core components						
High standards	3.71**	0.522	3.53–3.89	3.96	0.572	3.77–4.16
Rigorous curriculum	3.64**	0.484	3.47–3.80	3.89	0.605	3.68–4.09
Quality instruction	3.78	0.532	3.59–3.96	3.92	0.642	3.70–4.14
Culture of learning	3.80*	0.561	3.61–3.99	4.03	0.622	3.82–4.24
External community	3.43**	0.577	3.23–3.62	3.74	0.674	3.51–3.97
Performance accountability	3.59*	0.547	3.40–3.78	3.80	0.629	3.58–4.01
Key processes						
Planning	3.61***	0.492	3.44–3.78	3.89	0.578	3.69–4.09
Implementing	3.66**	0.485	3.49–3.82	3.92	0.579	3.72–4.16
Supporting	3.77	0.512	3.60–3.95	3.94	0.644	3.72–4.17
Advocating	3.65*	0.517	3.47–3.83	3.86	0.640	3.64–4.08
Communicating	3.66*	0.540	3.48–3.85	3.89	0.633	3.68–4.11
Monitoring	3.60**	0.505	3.42–3.77	3.83	0.591	3.63–4.04

$N = 35$; *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

themselves higher at time 2 than they did at time 1. For example, the *total effectiveness* score at time 1 is 3.66. At time 2, the *total effectiveness* score is 3.89.

Effect sizes based on Cohen's d (Cohen 1988) range from 0.25 for the core component of *quality of instruction* to 0.52 for the key process of *planning*. Additionally, the standard deviations are larger at time 2 than at time 1. While we did not expect the two administrations to be significantly different from each other, it is possible that after the first administration, principals are keyed into paying more attention to their learning-centered behaviors or they did improve their learning-centered behaviors. On the other hand, the first administration was a part of regular VAL-ED use, and the second administration was for our research purposes.

When comparing the time 1 scores with time 2 scores for the teachers' ratings of learning-centered behaviors, there are significant differences between the administrations of the VAL-ED. The one exception is for the core component of *external community* (see Table 3). For example, at time 1 on average, teachers rate their principals with a *total effectiveness* score of 3.67, whereas at time 2, teachers rate their principals with a *total effectiveness* score of 3.58. This is an effect size of 0.17.

Effect sizes range from a low of 0.07 for the core component of *external community* (time 1 = 3.55; time 2 = 3.51) to a high of 0.21 for the key process of *communicating* (time 1 = 3.72; time 2 = 3.61). Whereas the principals consistently rate themselves significantly higher at time 2 compared to time 1, teachers consistently rate principals significantly lower at time 2 compared to time 1. Perhaps, teachers are more aware of

Table 3 Teacher level means

	Time 1			Time 2		
	Mean	SD	95% Confidence intervals	Mean	SD	95% Confidence intervals
Total score	3.67**	0.477	3.55–3.78	3.58	0.563	3.45–3.71
Core components						
High standards	3.77***	0.470	3.66–3.89	3.68	0.554	3.55–3.81
Rigorous curriculum	3.63**	0.490	3.51–3.75	3.54	0.577	3.40–3.67
Quality instruction	3.71**	0.455	3.60–3.82	3.61	0.559	3.48–3.74
Culture of learning	3.74***	0.499	3.62–3.86	3.63	0.590	3.49–3.77
External community	3.55	0.522	3.42–3.67	3.51	0.562	3.37–3.64
Performance accountability	3.59**	0.516	3.47–3.71	3.50	0.611	3.35–3.64
Key processes						
Planning	3.64**	0.484	3.52–3.75	3.55	0.573	3.42–3.69
Implementing	3.64**	0.493	3.52–3.76	3.55	0.572	3.42–3.69
Supporting	3.69***	0.493	3.58–3.81	3.58	0.592	3.44–3.72
Advocating	3.66**	0.492	3.54–3.78	3.59	0.538	3.46–3.72
Communicating	3.72**	0.475	3.61–3.84	3.61	0.569	3.48–3.75
Monitoring	3.66*	0.465	3.55–3.77	3.60	0.574	3.46–3.73

$N = 71$;*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

the learning-centered behaviors that principals should be doing and are more inclined to note the absence of the behaviors. As with the principal ratings, the standard deviations are larger at time 2. Again, the time 2 administration was for our research purposes only.

It should be noted that consistently no matter the rater, principals are rated as being more than satisfactorily effective and close to being highly effective. In other words, despite the mean difference being significantly different, principals are seen as being effective in their learning-centered behaviors.

7 Discussion and conclusion

In examining the correlations between time 1 and time 2, we consistently find that the two administrations of the VAL-ED are positively correlated with large magnitudes in this particular sample of real users. This suggests that principals were rated similarly at both administrations of the VAL-ED. Yet, the comparison of means suggests that the time 1 and time 2 ratings are significantly different from each other when looking at the principal and teacher ratings with the scores being slightly higher at time 2 for principal self-ratings and slightly lower at time 2 for teacher ratings for these principals. While the correlation results are what we expected, the comparison of means was somewhat surprising.

In order to gain a deeper understanding of why time 1 and time 2 mean differences were significantly different, we looked at the score by comparing the performance categories that correspond to the score. The scores of the VAL-ED are divided into four performance levels. A panel of 22 educational experts including principals, teachers, supervisors, leadership researchers, and policymakers determined cut scores for four categories of principal effectiveness using the national field trial data, which consisted of 300 schools across the USA (Elliott et al. 2008; Cravens et al. 2013) (see Porter et al. 2010 for more information on the national field trial). If a score is equal to or above 4.00, the principal is considered distinguished. The cut score for proficient is 3.60, while the cut score for basic is 3.29. Scores below 3.29 are below basic (Elliott et al. 2008; Cravens et al. 2013).

When comparing the changes between time 1 and time 2 for principal self-ratings, at both time 1 and time 2 for the most part, principals rate themselves as proficient in the sample used for the test-retest study. Despite the increases in scores, the average principal stays in the proficient category. Of course, there are some category shifts for individual principals, but on average, principal self-ratings stay in the proficient category in our current sample.

For teacher ratings, there are some shifts in categories going from proficient to basic between time 1 and time 2 for the average teacher ratings of the principals in this study. These changes are for the total score, the core component of rigorous curriculum, and the key processes of planning, implementing, and advocating. We looked at the individual scores for each of these to get a sense of how many principals are changing categories according to the teacher ratings. Of those principals whom teachers rated as having a *total effectiveness* score of proficient (3.60 or above) at time 1, 20% were rated as having a *total effectiveness* score below proficient at time 2. For the core component of rigorous curriculum, 25% of the principals rated as proficient at time 1 were rated as not proficient at time 2. Twenty-two percent of the principals rated as proficient in time 1 on planning were rated as not proficient at time 2 by their teachers with about 14 and 15% for implementing and advocating, respectively.

It is possible the process of executing the test-retest design resulted in these shifts of ratings by keying the respondents to learning-centered behaviors. In other words, respondents may be "on the lookout" for these behaviors in the future. The principals may be more likely to notice behaviors that fit with the learning-centered behaviors or have actually increased their learning-centered behaviors. On the other hand, the teachers may be noticing when certain behaviors are not being done as they are more aware of the behaviors that they are rating.

Another limitation of this study is the time lag. While the original study design called for the two administrations to occur within a month's time in actuality, the lag ranged from 2 to 29 weeks. There was only one district that fell within our ideal four-week period. We conducted the analyses on a subset of schools where the two administrations were within 2 months of each other. The patterns that we report are consistent with the patterns we see in the restricted sample.

Evaluation of principals is important for school improvement (Goldring et al. 2009b), and states are increasingly including principal evaluations in addition to teacher evaluations. In 2009, Goldring et al., conducted a comprehensive review and

critique of the available principal evaluation tools. When examining existing tools, Author (Goldring et al. 2009b) also finds that there is a limited evaluation of learning-centered behaviors that have been shown to be important for student learning and student success such as behaviors related to a rigorous curriculum and high quality instruction. Additionally, research finds that the assessments that districts use are of varying quality (Condon and Clifford 2012). Condon and Clifford (2012) reviewed eight school leadership measures assessing their validity and reliability. While they found more instruments in their review, they focused on those that were rigorous. These eight instruments include the Change Facilitator Style Questionnaire (CFSQ), the Diagnostic Assessment of School and Principal Effectiveness, the Instructional Activity Questionnaire, the Leadership Practices Inventory (LPI), the Performance Review Analysis and Improvement System for Education (PRAISE), the Principal Instructional Management Rating Scale (PIMRS), the Principal Profile, and the VAL-ED. The instruments vary in terms of what aspects of the principalship they measure such as instructional leadership, leadership style, and change leaders. There is also variation in terms of the assessment and assessor. Some instruments use interviews, others use observation, and still others use questionnaires from teachers, supervisors, self, and students or combinations of those groups.

In their review, Condon and Clifford (2012) found that half of the instruments had poor reliability, and three out of eight had moderate reliability. The VAL-ED was the only instrument reviewed with high reliability. Additionally, only two of the instruments that Condon and Clifford (2012) reviewed had been developed since 2000, which may have implications for the validity of the instruments. With the increased emphasis on principal evaluation, it is important to have instruments that are valid and reliable.

Overall, this study provides evidence of test-retest reliability for VAL-ED for the seven districts in this sample. The scores from both administrations of the VAL-ED are correlated with each other with the correlation having a large magnitude, which aligns with the desired outcomes of test-retest reliability studies for a given population.

The findings of this study add to the accumulation of evidence on the psychometric properties of the VAL-ED. The testing standards require that high stake accountability decisions be based on multiple measures. Based on the validity and reliability studies published, the VAL-ED can be used in combination with other measures when evaluating principals.

Acknowledgements The research reported here was supported by the Institute of Education Sciences, US Department of Education through grant numbers R305A0803070, R305B100013-01, and R305E100008 of the US Department of Education. Please direct all correspondence to Elizabeth Covay Minor (eminor1@nl.edu).

Compliance with ethical standards

Conflict of interest The authors declare a potential conflict of interest (e.g., a financial relationship with the commercial organizations or products discussed in this article) as follows: The Vanderbilt Assessment of Leadership in Education (VAL-ED) instrument is authored by Drs. Porter, Murphy, Goldring, and Elliott and copyrighted by Vanderbilt University, all of whom receive a royalty from its sales by Discovery Education Assessment. The VAL-ED authors and their research partners have made every effort to be objective and data based in statements about the instrument and value the independent peer review process of their research. With any publication, readers should judge the facts and related materials for themselves.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington: American Educational Research Association.

Bryk, A. S., Sebring, P. B., Allensworth, E., Luppescu, S., & Easton, J. Q. (2010). *Organizing schools for improvement: lessons from Chicago*. Chicago: University of Chicago Press.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Carmine, E., & Zeller, R. (1979). *Reliability and validity assessment*. Thousand Oaks: Sage Publications, Inc..

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: Lawrence Erlbaum Associates, Inc..

Condon, C. & Clifford, M. (2012). Measuring principal performance: How rigorous are commonly used performance assessment instruments? Retrieved from http://www.air.org/sites/default/files/downloads/report/Measuring_Principal_Performance_0.pdf.

Covay Minor, E., Porter, A., Murphy, J., Goldring, E., Cravens, X., & Elliott, S. (2014). A known group analysis study of the vanderbilt assessment of leadership in education in US elementary and secondary schools. *Educational Assessment, Evaluation and Accountability, 26*, 29–48.

Cravens, X., Goldring, E., Porter, A., Polikoff, M., Murphy, J., & Elliott, S. (2013). Standard setting for principal leadership assessment: a deliberative process. *Educational Administration Quarterly, 49*(1), 124–160.

Elliott, S. N., Murphy, J., Goldring, E., & Porter, A. (2008). *VAL -ED users' guide*. Nashville: Discovery Education Assessments.

Fan, X. (2001). Statistical significance and effect size in education research: two sides of a coin. *The Journal of Educational Research, 94*(5), 275–282.

Goldring, E., Porter, A. C., Murphy, J., & Elliott, S. (2009a). Assessing learning-centered leadership: connections to research, professional standards, and current practice. *Leadership and Policy in Schools, 8*(1), 1–36.

Goldring, E., Cravens, X., Murphy, J., Elliott, S., Porter, A., & Carson, B. (2009b). The evaluation of principals: what and how do states and urban districts assess? *The Elementary School Journal, 110*(1), 19–32.

Goldring, E., Cravens, X., Murphy, J., Porter, A., & Elliott, S. (2015). The convergent and divergent validity of the Vanderbilt Assessment of Leadership in Education (VAL-ED): instructional leadership and emotional intelligence. *Journal of Educational Administration, 53*(2), 177–196.

Hallinger, P. (2011). A review of three decades of doctoral studies using the Principal Instructional Management Rating Scale: a lens on methodological progress in educational leadership. *Educational Administration Quarterly, 4*, 271–306.

Hallinger, P., & Murphy, J. (1985). Assessing the instructional management behavior of principals. *Elementary School Journal, 86*(2), 217–247.

Hallinger, P., Wang, W.-C., & Chen, C.-W. (2013). Assessing the measurement properties of the Principal Instructional Management Rating Scale: a meta-analysis of reliability studies. *Educational Administration Quarterly, 49*, 272–309.

Leithwood, K., & Montgomery, D. J. (1982). The role of the elementary school principal in program improvement. *Review of Educational Research., 52*(3), 309–339.

Leithwood, K., Jantzi, D., & McElheron-Hopkins, C. (2006, Feb). The development and testing of a school improvement model. *School Effectiveness and School Improvement, 17*(4), 441–464.

Leithwood, K., Louis, K. S., Anderson, S., & Wahlstrom, K. (2004). *Review of research: how leadership influences student learning*. Minneapolis: Center for Applied Research and Educational Improvement, University of Minnesota.

Leithwood, K., & Sun, J. (2012). The nature and effects of transformational school leadership: a meta-analytic review of unpublished research. *Educational Administration Quarterly, 48*(3), 387–423.

Marks, H. M., & Printy, S. M. (2003). Principal leadership and school performance: an integration of transformational and instructional leadership. *Educational Administration Quarterly, 39*(3), 370–397.

May, H., & Supovitz, J. A. (2011). The scope of principal efforts to improve instruction. *Educational Administration Quarterly., 47*(2), 332–352.

Murphy, J. (1989). Educational reform in the 1980s: explaining some surprising success. *Educational Evaluation and Policy Analysis, 11*(3), 209.

Murphy, J. (1992). School effectiveness and school restructuring: contributions to educational improvement. *School Effectiveness and School Improvement, 3*(2), 90–109.

Murphy, J., Hallinger, P., Weil, M., Mitman, A., & 3. (1983). Problems with research on educational leadership: issues to be addressed. *Educational Evaluation and Policy Analysis, 5*, 297–305.

Murphy, J., Hallinger, P., & Mesa, R. P. (1985). School effectiveness: checking progress and assumptions and developing a role for state and federal government. *Teachers College Record, 86*(4), 615–641.

Murphy, J., Elliott, S., Goldring, E., & Porter, A. C. (2007). Leadership for learning: a research-based model and taxonomy of behaviors. *School Leadership & Management, 27*(2), 179–201.

Murphy, J., Goldring, E., & Porter, A. C. (2012). *Building productive principal evaluation systems*. Nashville: Vanderbilt University, VALED.

National Instrument. (2006). Improving accuracy through averaging. Retrieved from http://www.ni.com/white-paper/3488/en/.

Petrides, K. V., Pérez-González, J. C., & Furnham, A. (2007). On the criterion and incremental validity of trait emotional intelligence. *Cognition and Emotion, 21*, 26–55.

Polikoff, M. S., May, H., Porter, A. C., Elliott, S. N., Goldring, E., & Murphy, J. (2009). An examination of differential item functioning on the Vanderbilt Assessment of Leadership in Education. *Journal of School Leadership, 19*(6), 661–679.

Porter, A. C., Murphy, J., Goldring, E., Elliott, S. N., Polikoff, M. S., & May, H. (2008). *VAL - ED technical manual*. Nashville: Vanderbilt University.

Porter, A. C., Polikoff, M. S., Goldring, E. B., Murphy, J., Elloitt, S. N., & May, H. (2010). Investigating the validity and reliability of the Vanderbilt Assessment of Leadership in Education. *The Elementary School Journal, 111*(2), 282–313.

Preston, C., Goldring, E., Guthrie, J., & Ramsey, R. (2012). *Conceptualizing essential components of effective high schools*. Nashville: Paper presented at the National Conference on Achieving Success at Scale: Research in Scaling Up Effective Schools.

Robinson, V. M. J., Lloyd, C. A., & Rowe, K. J. (2008). The impact of leadership on student outcomes: an analysis of the differential effects of leadership types. *Educational Administration Quarterly, 44*(5), 635–674.

Supovitz, J., Sirinides, P., & May, H. (2009). How principals and peers influence teaching and learning. *Educational Administration Quarterly, 46*(1), 31–56.