

Scaling of Educational Leadership Candidates' Commitment to National Standards: The ELCBS Scale

Dorothy C. Rea – Florida Gulf Coast University

Cecil F. Carter – Florida Gulf Coast University

Christopher M. Parfitt - Florida Gulf Coast University

Judy R. Wilkerson – Florida Gulf Coast University

Thomas C. Valesky – Florida Gulf Coast University

Abstract

Colleges of education are faced with Council for the Accreditation of Educator Preparation (2017) requirements to assess dispositions in addition to knowledge and skills. Preparation programs across the country are looking for ways to assess dispositions through valid and reliable measures. We describe the validation of a survey instrument to assess the dispositions of master's degree candidates in Educational Leadership. We used the dispositions outlined in the Educational Leadership Policy Standards developed by the Council of Chief State School Officers. These standards provide a viable content domain from which to assess leaders' affective learning. The instrument described in this article, the Educational Leader Candidate Belief Scale (ELCBS), was developed for measuring educational candidates' leadership dispositions.

Keywords: dispositions, standards, CAEP, CCSSO, Rasch

Scaling Educational Leadership Candidates' Commitment to National Standards: The ELCBS Scale

In this article, we describe the validation process for an instrument, Educational Leadership Candidate Belief Scale (ELCBS), which was originally designed to assess the level of commitment to Interstate School Leaders Licensure Consortium (ISLLC) dispositions of master's level educational leadership candidates. Assessing dispositions is important for a number of reasons, the most vital of which is to ensure pre-service preparation programs are graduating future school leaders who possess dispositions necessary for success. In addition, the Council for the Accreditation of Educator Preparation (CAEP) requires institutions to assess candidate dispositions (Council for the Accreditation of Educator Preparation, 2017). We began the process of developing an assessment instrument, the ELCBS, by using the dispositions enumerated in the document developed as a companion piece to the 2008 National Educational Leadership Policy Standards (Council of Chief State School Officers, 2008a) titled *Performance Expectations and Indicators for Education Leaders* (Council of Chief State School Officers, 2008b).

Dispositions Defined

To assess dispositions effectively, one needs to define the construct. Katz (1993) defined dispositions as patterns of behavior, exhibited frequently and intentionally in the absence of coercion, representing a habit of mind. Later, Perkins (1995) defined dispositions as the proclivities that lead us in one direction rather than another within the range of freedom possessed. Then in 2001, Ritchhart viewed dispositions as a collection of cognitive tendencies that capture one's patterns of thinking, addressing the gap between abilities and actions.

Damon (2007) warned that for certification-related assessment, dispositions "must be based on clearly defined principles rather than the fuzzy intuitions of whoever happens to be in charge of the process at any one time" (p. 368). Dottin (2009) concluded that educators are just beginning to grapple with the definition. He further stated, "Dispositions, therefore, concern not only what professional educators can do (ability), but also what they are actually likely to do (actions)" (p. 85). The plethora of definitions, then, is of concern. In our work with dispositions and in working with our students, we have adopted the definition used by Wilkerson and Lang (2007): dispositions are attitudes, values, and beliefs that influence the use of knowledge and skills. We prefer this definition for two reasons: it encompasses other authors' definitions and it focuses on the observable behaviors, which can be evidence of dispositions.

Disposition Assessments

In an exploratory, qualitative study, Lindahl (2009) examined if and how dispositions were taught and assessed in principal preparation programs. All respondents who were interviewed considered that dispositions were a key element of principal preparation. In almost all cases the dispositions identified in national standards were used. He concluded that if dispositions were to be addressed in educational leadership programs, a valid and reliable instrument should be developed. However, he qualified this conclusion with cautionary questions about the reliability of assessment practices:

1. Is it possible to develop an effective process for assessing dispositions, or are there some idiosyncratic elements that might not conform well to even a well thought-out process?
2. What levels of expectations ("dispositional tolerance") should be set and what levels define a passing score? Who determines this, and how?
3. How can evaluators prevent their personal biases in favor or against specific dispositions from entering into their subjective judgment of candidates?

4. Are dispositions synergistic in nature, where the whole is greater than a sum of the parts?

Currently at most institutions, the assessment of dispositions is largely dependent on the use of Likert scales of self-reported beliefs less closely linked to the standards than instruments that measure cognitive abilities. Examples of instruments measuring cognitive abilities are reported by Richardson and Onwuegbuzie (2003); Brown, King, and Herron (2008); and Schulte, Edwards, and Edick (2008). Scale development is typically based on locally developed construct definitions such as those identified above, rather than the standards directly. These studies also rely on classical statistical procedures, including descriptive statistics, factor analysis, and chi square tests.

Research Questions

The gap in the literature of leader dispositions assessment research is twofold. First, there is limited attention to building a scale that systematically samples from the content domain needed for accountability and accreditation (i.e., national standards). Second, the measurement process is largely reliant on statistics that fail to address the assumptions for their use and/or do not lead to research designs that take advantage of pairing dispositions results with interval level achievement scores. The questions explored here are:

1. Does the Educational Candidates Leadership Belief Scale (ELCBS) provide a valid and reliable measurement of master's level educational leadership students' commitment to national standards?
2. Do students who have had specific training in dispositions for educational leaders outperform students who have not been trained in these dispositions?

Methodology

In the ELCBS, the educational leadership candidates choose whether they agree or disagree on each of 53 statements, which we created. There are eight to 10 items on the ELCBS per each Performance Expectation (PE). Each statement was classified based on our expectation of how difficult the statement would be to answer correctly, with the goal of ensuring variability in responses. This is important because without variability in responses there is no measurement, only confirmation. We also associated each item in the instrument to the affective taxonomy by Krathwohl, Bloom, and Masia (1964) as adapted to educational leadership (Wilkerson & Lang, 2011). This adapted taxonomy includes the following steps to internalization of an attitude, value, or belief: *Unaware*; *Receiving*; *Responding*; *Valuing*; *Organization*; and *Characterizing*. Each successive step in the taxonomy increases internalization of the attitude, value, or belief. Through the application of the taxonomy to the instrument, we recognize that there are levels of internalization of attitudes, values, beliefs and through internalization, an individual accepts or conforms to that attitude, value, or belief, resulting in direct behavior.

Existing measures, such as the one proposed by Brown, King, and Herron (2008), showing virtually no variability, are less likely to explain differences in performance. To avoid having no variability of responses in which respondents simply agreed to all items without much thought, we developed items so there was a mix of dichotomous responses of "agree" and "disagree." In other words, we worded the items so respondents must think carefully about each item before answering "agree" or "disagree."

After administering the ELCBS to a sample of undergraduates and graduate students described below, we then tested the scale using the Rasch model of item-response theory. The actual item scores were used to redefine the theory related to the construct of attitude toward the Standards. This definition of the construct can then be used to determine decisions about program design.

Instrument and Measurement Method

Thurstone (1928) defined attitudes to include a person's inclinations, feelings, prejudice, bias, preconceived notions, ideas, fears, threats, and convictions about any topic. Both Likert and Thurstone scales are comprised of statements to which respondents agree or disagree, but the Thurstone's technique requires a dichotomous decision (agree/disagree only), while Likert provides for a rating scale, typically five-points, from strongly agree to strongly disagree with a neutral midpoint. Roberts, Laughlin, and Wedel (1999) examined the relationship between Likert and Thurstone agreement scaling, recommending the Thurstone scale when extreme positions (e.g., high/low levels of commitment) are of interest:

...the Likert procedure may falter for individuals who hold extreme attitudinal positions when responses result from some type of ideal point process. This is because the Likert procedure is functionally a cumulative model of the response process, and as such, it is not always compatible with responses from an ideal point process. In contrast, the Thurstone procedure is functionally an unfolding model, and thus, it does correspond to the situation in which responses follow from an ideal point process. Due to this correspondence, the Thurstone procedure does not suffer from the degraded validity exhibited with the Likert method when individuals with extreme attitudes are measured. (pp. 229-230)

The Rasch (1960) model is the simplest form of item-response theory, calling for careful delineation of the construct during the design stage (Wilson, 2005). Conceptually, the idea behind the Rasch model is simple. The ability (or, in this case, commitment) of individuals and the difficulty of items influence each other conjointly. The Rasch (1960) model places them on the same interval scale, so predictions about one from the other can be made. Rasch established the mathematical relationship between a person's ability (or commitment) and the difficulty of an item, demonstrating that the probability of providing a correct response was related to the ability (or commitment) of the respondent.

Item response theory and the family of Rasch models permit ordinal level data, including dichotomous and rating scale items, to be converted to an interval scale. This allows more appropriate use of common statistics, providing advantages over a simple raw score (count) of correct responses. With a purposive sample and a skewed distribution, inferential statistics are not appropriate. Rasch modeling is sample independent and requires neither a large sample nor a normal distribution (Bond & Fox, 2007). Rasch allows the user to create an interval level scale that can then be used for associational or intervention research designs in subsequent studies. Validity and reliability statistics can also be reported (Linacre, 2003). Rasch is extensively used by most modern test publishers, such as Pearson, in the development of major high-stakes tests.

Sample

Four groups of students were assessed. One group was composed of master's level educational leadership candidates (n=77), for whom the instrument was designed (please refer to research question number 1). A second group was composed of doctoral level students (n=25). In addition, there were two groups of undergraduate students: one group was composed of sophomores enrolled in an instructional technology class (n=24) and one group of seniors enrolled in an undergraduate measurement course (n=48). The four groups completed the instrument to allow for comparisons across three levels of university experience. The total sample size was 174, as shown in Table 3.

Data Analysis

Statistical Results on ELCBS

Descriptive statistics are provided in Table 1. Note that these results include both raw scores for persons and scaled scores for both persons and items. In Rasch measurement, the extent to which items and people fit the mathematical model are reported for outfit and infit, which differ based on the extent to which outliers are incorporated. The expected mean square fit statistic is 1.0 with a standardized z value of zero. For ELCBS, the fit statistics are provided in Table 2. Note that both items and persons fit the model well, meaning that scores were mathematically predictable based on the relative difficulty of each item and the relative commitment level of each individual.

Table 1.

Descriptive Statistics for ELCBS

Statistic	Person Raw Scores	Rasch Person Measure (Commitment)	Rasch Item Measure (Difficulty)
Mean	30.4	60.97	50.00
Standard Deviation	5.4	7.87	13.69
Maximum	44.0	96.58	84.51
Minimum	7.0	44.32	22.67

Table 2.

Fit Statistics for ELCBS

Statistic	Infit		Outfit	
	Mean Square	Standardized Z	Mean Square	Standardized Z
Persons	1.0	.0	.99	.0
Items	.99	.0	.99	.1

Figure 1 provides the variable map from the Rasch model, showing the relationship between items and people on the same vertical scale. Note that the person distribution on the left side of the map varies from about 45 to 75 correctly answered items, and scores are skewed toward the higher end of the scale. As research continues with this instrument and students from lower levels of experience are added to the scale, we expect the distribution will be more normal. The right side of the map displays item difficulty with easy items at the bottom and harder items at the top. This side of the map shows there are still items at the bottom right that are too easy for these students. The spread of items on the right indicates that there is good coverage of the construct however, supporting construct validity of the instrument.

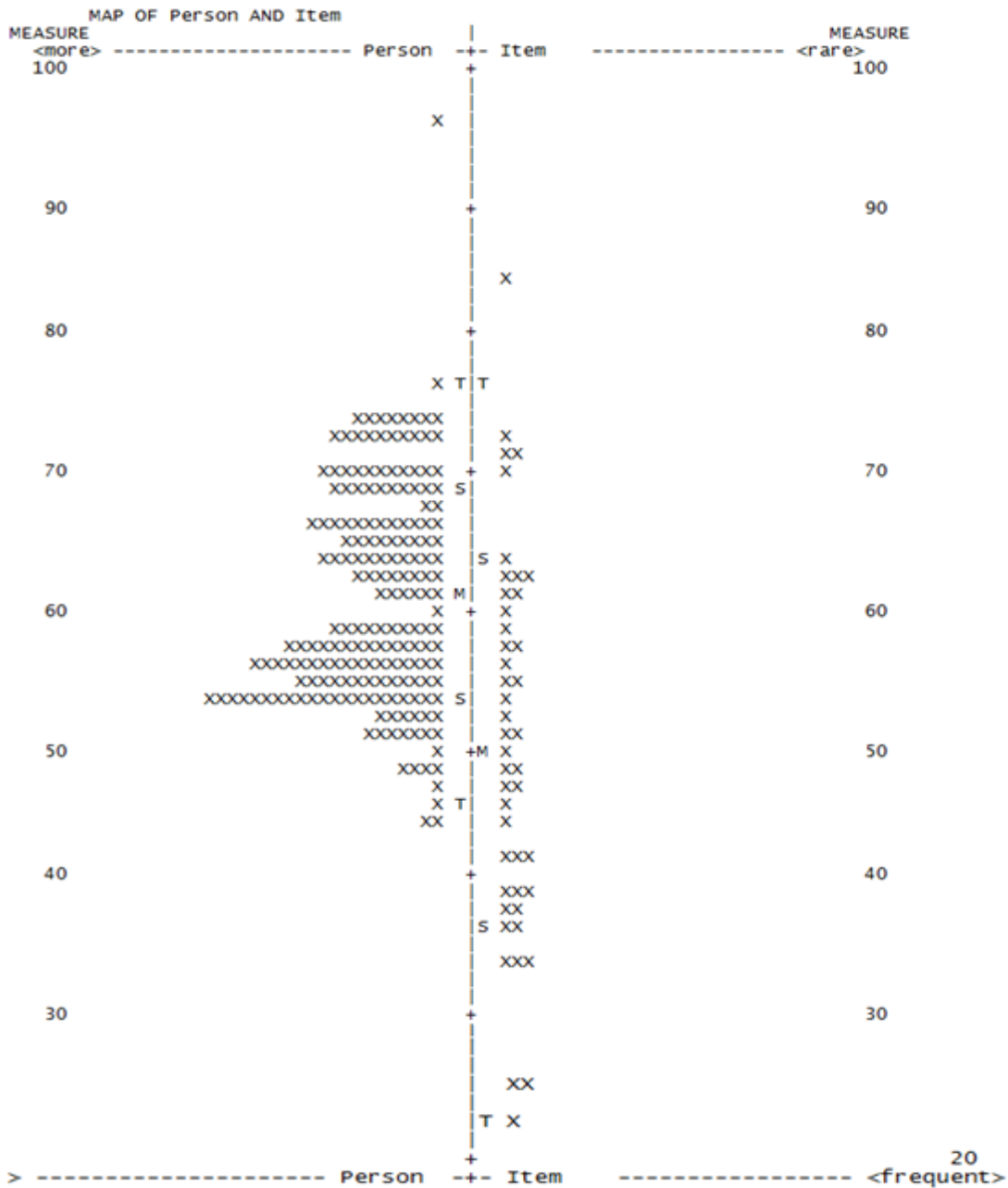


Figure 1. ELCBS variable map

In Figure 2, the item measures (in measure order from most difficult at the top to least difficult items at the bottom) are provided for each item on the instrument except the eight which were dropped from the analysis—these eight items were not functioning well from a mathematical perspective in that their point biserial correlations were negative. Two items remaining in the ELCBS have a negative point biserial, but they were left in the scale since they tapped important traits and did not have a noticeable difference in overall scores. The expected range of .5 to 1.5 for infit and outfit statistics (Linacre, 2003) held for each item. Although there were two items with a negative point biserial correlation, each of the remaining items is functioning well.

Reliability and separation statistics are acceptable, with Cronbach's alpha (KR-20) estimated at .72. The person reliability of .70 indicates satisfactory separation of about one and a half levels (separation = 1.61). Item reliability and separation are excellent at .97 and 6.07. The higher separation index for items shows that there is more variability in the scores, including a larger standard deviation. The lower index for persons indicates that the population is relatively homogeneous.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S. E.	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD	PT-MEASURE CORR.	EXP.	EXACT OBS%	MATCH EXP%	Item
10	20	186	84.51	2.46	.95	-.2	.94	-.1	.31	.25	89.8	89.5	10.1.children can learn
3	47	187	73.19	1.78	1.05	.5	1.13	1.0	.25	.32	76.5	75.7	03.1.resources & opps
8	50	184	71.96	1.75	1.14	1.7	1.23	1.9	.13	.33	70.1	74.1	08.1.vision & environ.
2	52	186	71.54	1.73	1.05	.6	1.10	.9	.26	.33	75.3	73.5	02.1.consensus
42	57	186	70.08	1.69	.85	-2.1	.79	-2.2	.52	.34	75.8	71.7	42.5.talk before reporting
44	79	185	64.07	1.59	.74	-5.3	.75	-4.0	.65	.34	85.4	66.3	44.5.reflection worth it
19	82	183	63.22	1.58	1.12	2.3	1.16	2.3	.19	.34	60.1	65.7	19.2.research/data help
41	84	185	62.83	1.57	.94	-1.2	.96	-.6	.41	.34	70.3	65.7	41.5.report bad princ.
16	88	186	62.14	1.57	1.04	.8	1.04	.7	.29	.34	62.9	65.3	16.2.stats lie
30	90	186	61.50	1.56	1.29	5.4	1.34	4.8	-.02	.34	45.2	65.2	30.4.sch pers know best
53	93	187	60.96	1.56	.90	-2.1	.89	-1.7	.46	.34	74.3	65.1	53.6.join civic orgs.
28	98	185	59.50	1.57	1.07	1.4	1.09	1.3	.25	.34	61.6	64.6	28.3.princ.help pers.
4	104	187	58.28	1.57	1.09	1.8	1.34	4.4	.19	.34	60.4	64.7	04.1.vision buy-in
15	105	183	57.41	1.59	.97	-.6	.95	-.6	.37	.33	66.1	64.7	15.2.degree sufficies
34	104	181	57.18	1.60	1.06	1.1	1.07	.9	.26	.33	62.4	64.7	34.4.princ. reach out
40	112	185	56.09	1.59	.88	-2.3	.83	-2.2	.48	.33	66.5	65.6	40.5.disc.beh.plan parents
36	116	186	55.07	1.60	1.10	1.7	1.15	1.7	.19	.32	64.5	66.3	36.4.princ. research comm.
11	119	185	54.22	1.62	1.07	1.2	1.09	1.0	.23	.32	65.4	67.1	11.2.workplace prep
7	121	185	53.65	1.63	1.00	-.1	1.00	.0	.32	.32	65.9	67.9	07.1.change & evidence
24	127	186	52.30	1.66	.91	-1.4	.83	-1.7	.43	.31	72.0	69.8	24.3.don't delegate
48	131	187	51.29	1.68	.94	-.9	.86	-1.2	.39	.30	72.2	71.2	48.6.gen.couns.expert
14	131	186	51.18	1.69	.90	-1.5	.81	-1.7	.44	.30	73.1	71.4	14.2.work alone
37	138	186	49.12	1.75	.90	-1.2	.82	-1.4	.42	.28	75.8	74.6	37.4.princ.relat.resources
49	138	186	48.96	1.75	.98	-.2	.92	-.5	.32	.29	73.1	74.7	49.6.circumvent law if nec.
17	139	185	48.54	1.78	.95	-.5	.96	-.2	.33	.28	77.8	75.4	17.2.data takes time
43	144	187	47.34	1.81	.84	-1.9	.67	-2.5	.51	.28	78.1	77.2	43.5.degree is all needed
26	145	186	46.86	1.84	1.07	.7	1.13	.9	.17	.27	77.4	78.1	26.3.parents write disc.
38	147	183	45.30	1.93	1.07	.7	1.13	.8	.15	.26	79.8	80.3	38.5.report uneth.tchr.
39	152	185	43.97	1.99	1.18	1.5	1.37	1.9	-.01	.25	82.2	82.2	39.5.hire relative
22	157	186	42.28	2.09	1.13	1.0	1.56	2.4	.00	.23	84.4	84.4	22.3.resources for need
23	157	186	42.28	2.09	.95	-.3	1.01	.1	.28	.23	84.4	84.4	23.3.pay high perf.more
46	159	187	41.66	2.11	.98	-.1	.86	-.6	.28	.23	85.0	85.0	46.6.equity not possible
21	162	185	39.54	2.29	.93	-.4	1.09	.5	.27	.21	87.6	87.6	21.3.teachers choose sch.
12	163	186	39.43	2.29	1.00	.0	1.01	.1	.20	.21	87.6	87.6	12.2.diversity a problem
32	164	186	38.73	2.33	.97	-.1	.99	.1	.24	.21	88.2	88.2	32.4.email parents
35	165	186	38.18	2.37	.96	-.2	.85	-.5	.27	.20	88.7	88.7	35.4.SACS need diversity
1	167	187	37.59	2.42	.96	-.2	.90	-.3	.26	.20	89.3	89.3	01.1.single data
5	168	187	36.99	2.48	1.05	-.3	1.21	.8	.11	.20	89.8	89.8	05.1.vision & impr.plans
33	167	185	36.45	2.54	.99	.0	.98	.0	.20	.19	90.3	90.3	33.4.parents want control
27	168	183	34.60	2.74	1.02	-.2	1.18	.7	.14	.18	91.8	91.8	27.3.low bid works
9	171	186	34.47	2.74	.97	-.1	.90	-.2	.22	.17	91.9	91.9	09.1.principal's vision
20	171	185	33.79	2.83	.95	-.2	.92	-.1	.23	.17	92.4	92.4	20.2.don't mainstream
13	180	186	24.55	4.18	.94	.0	.51	-1.0	.25	.11	96.8	96.8	13.2.sameeness helps
25	180	186	24.55	4.18	.95	.0	.75	-.4	.20	.11	96.8	96.8	25.3.don't fix if ok
18	180	185	22.67	4.56	.96	.0	.51	-.9	.22	.10	97.3	97.3	18.2.indivs.take time
MEAN	126.5	185.5	50.00	2.08	.99	.0	.99	.1			77.8	77.8	
S. D.	41.2	1.3	13.69	.70	.10	1.5	.21	1.6			11.8	10.8	

Figure 2. Items in Measure Order

It was hypothesized that students who were taught about the educational leadership dispositions would outperform students without such instruction. Descriptive statistics for all groups are listed in Table 3. From this data, we conducted an Analysis of Variance (ANOVA) in SPSS, resulting in an F value of 78.245 at 173 degrees of freedom, with $p=.000$. This analysis indicated a significant difference between the groups we had defined (sophomores, seniors, master's, and doctoral-level students). Since the difference in the mean ELCBS scores was significant, a post-hoc Least Significant Difference (LSD) test was conducted. The result of the LSD test indicated a significant difference, again with $p=.000$, between all pairs of groups except sophomores and seniors ($p=.315$). The group mean difference between master's students (the group being trained in dispositions) is about 1.8 standard deviations higher than seniors (with little to no experience in schools) and about 1.1 standard deviations higher than doctoral

students. One might believe that the doctoral group would have been highest, but they have not yet benefited from instruction in dispositions.

Table 3.
SPSS Output of Descriptive Statistics for ELCBS

Group	Number	Mean	Standard Deviation
Sophomores	24	53.50	3.95
Seniors	48	54.72	3.45
Master's	77	66.40	5.90
Doctoral	25	59.24	4.13
Total	177	60.37	7.40

Conclusions and Implications

Evidence of construct validity and reliability (internal consistency and separation reliability) were presented in this study. Use of the ELCBS is, therefore, providing useful data to document educational leadership program strategies that are working. This is evidence of construct validity.

In most contexts, we would expect doctoral students to score higher than master's level students. In the research reported in this study, however, only master's students are receiving training in dispositions, so we view this as a positive finding, supporting our general hypothesis that training can make a difference in the development of positive dispositions in educational leadership. The master's level students were trained during their program to improve their dispositions in educational leadership. This is not the case with the doctoral students since this cohort of doctoral students was new to the program and the cohort was composed of only a few students who completed a master's degree in Educational Leadership, but had many from curriculum and instruction, and an equal number from higher education with degrees outside education.

Eight items were dropped from the analysis as items that were not working well mathematically as described above. These items will need to be reviewed. In addition, all items will need to be reviewed for placement on Krathwohl, Bloom, and Masia (1964) taxonomy as adapted by Wilkerson and Lang (2011).

Our goal is to increase students' awareness of the proper dispositions to be successful school leaders. Several possibilities for improving candidate dispositions in this area exist. Coursework that encompasses the Performance Expectations could be added into the program's curriculum. Candidates could then articulate their positions to appropriate audiences. Candidates could be counseled on an individual basis concerning dispositional areas in which to improve. A third possibility for improving dispositions would be to include special projects related to Performance Expectations in the candidate internships. This option might be especially viable; inasmuch as internships includes experiences related to social justice.

References

- Bond, T., & Fox, C. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Brown, S., King, G., & Herron, S. (2008). Assessing dispositions of administrative intern candidates. *AASA Journal of Scholarship & Practice*, 5(1), 27-34.
- Council for the Accreditation of Educator Preparation. (2017). *Standards for advanced programs*. Retrieved from <http://caepnet.org/standards/standards-advanced-programs>
- Council of Chief State School Officers. (2008a). *Educational leadership policy standards*. Washington, DC: Author.
- Council of Chief State School Officers. (2008b). *Performance expectations and indicators for educational leaders*. Washington, DC: Author.
- Damon, W. (2007). Dispositions and teacher assessment: The need for a more rigorous definition. *Journal of Teacher Education*, 58(5), 365-369.
doi:10.1177/0022487107308732
- Dottin, E. (2009). Professional judgment and dispositions in teacher education. *Teaching and Teacher Education: An International Journal of Research and Studies*, 25(1), 83-88.
- Katz, L. G. (1993). *Dispositions: definitions and implications for early childhood practice*. Retrieved from <http://ceep.crc.uiuc.edu/eecearchive/books/disposit.html>
- Krathwohl, D., Bloom, B., & Masia, B.B. (1964). *Taxonomy of educational objectives, Book 2: Affective domain*. New York, NY: Longman.
- Linacre, J. M. (2003). *A user's guide to WINSTEPS: Rasch-model computer programs*. Chicago, IL: MESA Press.
- Lindahl, R. (2009). Teaching in and assessing dispositions principal-preparations programs: A conundrum. In C. M. Achilles, B. J. Irby, B. Alford, & G. Perreault (Eds.), *Remembering our mission: Making education and schools better for students* (pp. 15-38). Lancaster, PA: Pro-active Publications.
- Perkins, D. (1995). *Outsmarting IQ: The emerging science of learnable intelligence*. New York, NY: The Free Press.
- Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research) (Expanded ed.). Chicago, IL: The University of Chicago Press.
- Richardson, D., & Onwuegbuzie, A. (2003). *Attitudes toward dispositions related to the teaching of pre-service teachers, in-service teachers, administrators, and college/university professors*. Retrieved from ERIC Document Reproduction Service (ED482689)
- Ritchhart, R. (2001). From IQ to IC: A dispositional view of intelligence. *Roper Review*, 23(3), 143-50.
- Roberts, J. S., Laughlin, J. E., & Wedel, D. H. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement*, 59(2), 211-233.
- Schulte, L., Edwards, S., & Edick, N. (2008). The development and validation of the Diversity Dispositions Index. *AASA Journal of Scholarship & Practice*, 5(3), 11-19.
- Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, 33, 529-554.
- Wilkerson, J. R., & Lang, W. S. (2007). *Assessing teacher dispositions: Five standards-based steps to valid measurement using the DAATS model*. Thousand Oaks, CA: Corwin Press.

Wilkerson, J. R., & Lang, W. S. (2011). Standards-based teacher dispositions as a necessary and measurable construct. *The International Journal of Educational and Psychological Assessment*, 7(2), 34-55.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.