Exploratory Analysis of Teacher Artifacts as Evidence of Educator Effectiveness Implementation Fidelity

Pete Goldschmidt, California State University Northridge
Alia Congdon, no affiliation

Abstract: We collected artifacts from 42 teachers participating in a statewide educator effectiveness system to examine the fidelity with which the formative components of the system were implemented. Specifically, we collected written feedback from principals to teachers and teacher professional growth goals. We developed indicators of quality for each and examined whether there were relationships between these two indicators as well as with observations. Overall, principal feedback was often aligned with observation scores and the quality was directly related to the number of observation elements scored. Feedback is readily partitioned into two constructs: *clarity of communication* and *instructional practices*. Feedback consistently demonstrated clarity of communication, but was less likely to address instructional practices. Importantly, novice teachers received poorer quality feedback than experienced teachers. Teacher Professional Growth Goals tended to be superficial and rarely included details such as specific action steps or measurable outcomes. Although exploratory, evidence that both feedback and growth goals varied to some extent by school implies that both feedback and growth goals can be impacted by better guidance.

Keywords: feedback, Professional Growth Goals, Educator Effectiveness, implementation fidelity

Teacher supervision and evaluation through the implementation of educator effectiveness systems has emerged as a significant resource that districts and states are using to improve instructional practices and, by extension, student academic outcomes. Noteworthy contributors to this line of reasoning are the increased consensus that teachers matter (Chetty, Friedman, & Rockoff, 2014a, b; Kane, McCaffrey, Miller, & Staiger, 2013; Nye, Konstantopoulos, & Hedges, 2004; Rivkin, Hanushek, & Kain, 2005). Historically formal teacher ratings found virtually all teachers equally effective (Weisberg, Sexton, Mulhern, & Keeling, 2009), and results from the Measures of Effective Teaching (MET) indicate variation in teaching effectiveness can be meaningfully measured (Kane et al., 2013). States and districts, incentivized by Race to the Top funding and flexibility from *No Child Left Behind* (NLCB, 2001), embarked on policies emphasizing educator effectiveness systems (EESs) that included both summative and formative indicators of effective teaching (Bell et al., 2012; Shepard, 2012; Steinberg & Donaldson, 2016). Indicators of student learning provide summative evidence of teacher effectiveness, while indicators of instructional and professional practices provide an opportunity for a formative process through principal[1] feedback and teacher professional growth plans.

An unintended consequence of new EESs is that they create tension between a principal's supervisory roles and formal summative evaluation processes that place higher stakes on observation and blur the line between supervision and evaluation (Marshall, 2013). While not explicitly stated, EESs create a greater emphasis on a principal being an instructional leader–– precisely the area that principals feel the greatest need for mentoring and professional development (Johnston, Kaufman, & Thompson, 2016). While some argue that formal feedback is an ineffective means of improving instructional quality (DuFour & Marzano, 2009; Marshall, 2013), evidence also suggests that principals matter (Branch, Hanushek, & Rivkin, 2012) and that the impact is generally indirect, though it includes instructional coaching (Hallinger & Heck, 1996).

Consistent with results indicating that reporting summary scores alone is less effective in improving performance than summary scores with feedback (Hattie & Timperley, 2007; Rose & Farrel, 2002), early indicators of EES outcomes imply that summative results alone have had minimal effectiveness (Shepard, 2012) and that the mechanism for success rests to some extent on formative processes (Taylor & Tyler, 2012). This result is not surprising given several decades worth of evidence indicating that supervision and evaluation of instruction is an important element of effective principalship (Hallinger & Murphy, 1985). Although researchers cite good feedback as an important component of supervision and evaluation (Hallinger & Murphy, 1985; Marshall, 2013), there has been no research on the quality of feedback provided to teachers in the current EES environment. Given the renewed emphasis on instructional leadership as a part of supervision and evaluation, we conducted an exploratory study examining the implementation fidelity of the two formative components of an EES: written feedback to teachers and teacher goal setting.

Although feedback is more productive when decoupled from evaluation (DuFour & Marzano, 2009; Marshall, 2013; Meyer, 1991), feedback is a mandatory component of EESs and is intended as an intervention to guide improved performance. Feedback and Professional Growth Plans (PGPs[2]) purport to be the formative mechanism through which EESs affect changes in teacher and subsequently student outcomes. Meta-analyses of the impact of feedback generally demonstrate positive effects on outcomes (Hysong, 2009; Kluger & DeNisi, 1996), that results are quite varied (Kluger & DeNisi, 1996), and that the content of feedback matters (Cianci, Seijts, & Klein, 2010; Hysong, 2006; Ilgen, Fisher, & Taylor, 1979; Larson, Patel, Evans, & Saiman, 2013; Smither and Walker, 2004). Arguments have been presented that informal feedback as part of supervision is more effective than formal written feedback (DuFour & Marzano, 2009; Marshall, 2013), but this supposition has not been systematically tested.

Feedback as a mechanism to improve performance depends on several factors: the source of the feedback, the composition of the feedback, and the recipient of the feedback. Although much of the evidence related to feedback implies its use is somewhat atheoretical (Larson, Patel, Evans, & Saiman, 2013), evidence related to new EESs suggests that, minimally, a systematized feedback process has gained traction such that it provides more regular feedback (Heneman & Milanowski, 2009). In practice, state systems require some sort of feedback to teachers either after an observation, at the end of the school year, or both[3]. Moreover, state EESs also require reflection or PGP, which ought to be directly linked to feedback (Danielson, 1996; Hattie & Timperley, 2007; Ilgen et al., 1979). This is explicitly detailed in the Framework For Teaching (FFT; Danielson, 1996) that is the basis for the teacher practices portion of EESs in many states.

Using artifacts collected from teachers, our goals were to provide preliminary evidence related to the fidelity of implementing feedback and PGPs and to prompt further, more systematic research into this line of evaluation. We focused on describing specific attributes of feedback and PGPs, but we did not attempt to test a specific theory[4] to potentially identify moderating factors related to how teachers respond to feedback. We developed two indicators based on the relevant literature: a Feedback Quality Indicator (FQI) and a Professional Growth Goal Indicator (PGGI). We used these artifacts to address three primary areas: the quality of the feedback that teachers are receiving, the quality of the PGPs that teachers are writing, and the relationships among the measures and observation scores.

Background on EES and Formative Processes

Many states and districts have developed theories of action that link improved student outcomes to teachers through an EES that identifies a spectrum of teacher effectiveness and allows for various interventions along the effectiveness continuum. Although teacher evaluation has existed in some form for some time, results often suffered from lack of face validity and were considered inconsequential (Stiggins & Duke, 1988) and of insufficient quality to yield reliable indicators of teacher performance––in that 98% of teachers were deemed (equally) effective in virtually every state (Weisberg et al., 2009). This literature review focuses on providing support for considering feedback and PGP quality, along with what quality might look like. Specifically, the review includes the impact of effective teaching, feedback as an improvement mechanism, feedback effects, and elements of effective feedback and PGPs.

The Impact of Effective Teaching

The underlying theory of action is that teacher instructional practices impact student outcomes and that teacher instructional practices are malleable through evaluation and feedback. There is reasonable consensus that teacher effectiveness varies among teachers (Darling-Hammond, 2004) and that teachers contribute meaningfully to student outcomes in terms of student achievement. Teacher effects vary from about 0.1

to about 0.5 *SDs* (Rivkin et al., 2005; Nye, Konstantopoulos, & Hedges, 2004). Consistent with results identifying teacher effects in general is evidence indicating that specific teacher practices impact student outcomes (Black & William, 1988; Hattie & Timperley, 2007). Importantly, emerging evidence indicates that teaching practices, as identified through observation (Kane & Staiger, 2012; Kane, Taylor, Tyler, & Wooten, 2011; Taylor and Tyler, 2012), are related to student outcomes as well. It is important to note that principals demonstrate meaningful effects on student outcomes similar in magnitude to teacher effects and that principal effects are both direct and indirect (Hallinger & Heck, 1996; Johnston, Kaufman, & Thompson, 2016).

Although limited, recent evidence suggests that evaluation systems can have a positive impact on student outcomes (Kane, Taylor, Tyler, & Wooten, 2011; Kimball et al., 2008; Milanowski, 2004; Taylor & Tyler, 2012). A key component is the fidelity with which principals can manage an evaluation system and provide meaningful guidance to teachers (Milanowski, 2004). This is consistent with previous findings suggesting that poorly conducted evaluations can have a negative impact on performance––in that the person evaluated is less clear about their performance because the formal evaluation does not match signals provided by the evaluator (Marshall, 2013; Meyer, 1991). In fact, Marshall (2013) suggests continuous supervision through mini-observations with informal feedback, which then forms the basis for the formal evaluation and feedback. In this way there are no surprises because the teacher has been apprised of performance throughout the year (Marshall, 2013). Goal setting is an important step linked to actualizing continued successful performance, closing performance gaps, and implementing feedback (Ilgen et al., 1979; Kinicki, Wu, Prussia, & McKee-Ryan, 2004). It is important to examine goal setting in conjunction with feedback.

Feedback as an Improvement Mechanism

Individual feedback can either be derived from the task, from oneself, or from a rater (Kluger & DeNisi, 1996). In this analysis, we focused on feedback derived from a rater (i.e., a principal). The potential impact of feedback is governed by the credibility of the rater, where credibility is related to both expertise and trustworthiness (Ilgen et al., 1979). How the feedback is communicated is also important in engendering acceptance by the recipient (Ilgen et al., 1979; Kinicki et al., 2004). Generally, teachers want instructional feedback (Marshall, 2013). In order for the recipient to accept the feedback, it must be accurate, consistent, and task-focused (Ilgen et al., 1979; Kinicki et al., 2004; Smither & Walker, 2004). Recipients are more likely to deem feedback as accurate if it is positive and aligns with self-perceived evaluations of the task. For teachers, this inclination is exacerbated by discrepancies between ratings and feedback. More affirmative ratings received on the observation rubric allow teachers to discount feedback, whereas alignment between the two is positively related with the desire to respond (Kinicki et al., 2004). Feedback should be clear about the gap between standards and performance (Larson et al., 2013). The more feedback that is directed at the task (as opposed to the recipient), the more likely a recipient is to accept the feedback as accurate (Kinicki et al., 2004; Larson et al., 2013). Consistent ongoing feedback is part of supervision and can help inform end of year written feedback, which helps summarize and codify important goals (Marshall, 2013). Response then takes the form of developing goals that are aligned with feedback (Ilgen et al., 1979).

Feedback potentially provides two mechanisms through which instructional practices may change. In one, feedback enhances reflection (McDonald & Boud, 2003; Winne & Butler, 1994), as feedback equips teachers to self-assess with additional information (Sadler, 1989)[5] and provides opportunities to reflect on goals and strategies, which is effective in enhancing outcomes (McDonald & Boud, 2003). In another, feedback provides a concrete mechanism through which teachers can be apprised of standards (Bell et al., 2012) and the gap between their practices and those standards/expectations (Hattie & Timperley, 2007; Ilgen et al., 1979; Kinicki et al., 2004).

Feedback Effects

Analyses of the impact of feedback demonstrate highly varied but positive effects (Hysong, 2009; Kluger & DeNisi, 1996) with an average effect size of approximately 0.4 (Kluger & DeNisi, 1996). Written corrective feedback is better than simply providing summary scores (Hattie & Timperley, 2007) and written feedback is given greater attention than scores alone (Rose & Farrel, 2002). Consistent with the notion of providing more than summary scores is the result that active (formative) evaluation is more useful and effective than passive evaluation (Black & William, 1998) and leads to increased use of effective teaching strategies (Scheeler, Dochy & Janssens, 2004).

Importantly, evidence is emerging that observed teacher practices, which form the basis of feedback, impact student academic performance. Kane et al. (2011) found that observed teacher practices are asso-

ciated with improved student performance. Effect sizes are approximately 0.14 *SD* (Kane et al., 2011). Similarly, Taylor & Tyler (2012) found effect sizes ranging from .064 *SD* in the year of evaluation to .112 *SD*, .158 *SD*, and .161 *SD* one, two, and three years after the evaluation, respectively. The effect sizes are based on student academic progress, and the observations used by Taylor & Tyler (2012) are based on the FFT (Danielson, 1996). Other observation rubrics demonstrate positive relationships with student learning outcomes consistent with the FFT (Kane & Staiger, 2012).

Elements of Effective Feedback

Feedback potential is enhanced in a feedback-rich environment, which affects the perceived accuracy of feedback (Kinicki et al., 2004). Three elements of feedback are generally taken into consideration: timing, credibility, and utility (Ilgen et al, 1979; Kimball, 2002; Kinicki et al., 2004; Larson et al., 2013; Smither & Walker, 2004). Utility is further refined into non-corrective feedback and corrective feedback. Non-corrective feedback, such as praise, has little impact because it carries little information about the task (Hattie & Timperley, 2007), and corrective feedback improves behavior (Scheeler et al., 2004). It is important to note that feedback recipients view feedback as more accurate when it is positive (Kinicki et al., 2004; Smither & Walker, 2004). Specific facets of corrective feedback have also been identified: directed and specific, detailed, immediate (Nicol & Macfarlane-Dick, 2006), and focusing on gaps between performance and expected performance or standards (Larson et al., 2013). Consistent with expectations, feedback that is directed to the task, clear (Black & William, 1998; Hattie & Timperley, 2007; Kinicki et al., 2004; Smither & Walker, 2004), and is specific (Scheeler et al., 2004) improves behavior. Feedback that identifies the type and the extent of error and is specific in ways to correct it is most effective (Larson et al., 2013; Scheeler et al., 2004). This includes clearly delineating goals, expectations, performance towards meeting those expectations, gaps in performance, and steps to close performance gaps (Hattie & Timperley, 2007; Heneman & Milanowski, 2004; Kimball & Milanowski, 2009; Larson et al., 2013; Morey, 2003; Scheeler, Dochy, & Janssens, 2004; Thurlings, Vermeulen, Kreijins, Bastiaens, & Stijnen, 2012; White, 2009). Evidence suggests that the more complex the task, the more specific feedback needs to be in order to improve outcomes (Eisner, 1992).

It is also clear that feedback provides guidance towards self assessment and reflection (Hattie & Timperley, 2007; Ilgen et al., 1979), and feedback to teachers is, in part, intended to facilitate self-reflection (Danielson & McGreal, 2000). Consistent with feedback, reflection and goals should be structured and consist of elements such as identifying strengths and weaknesses, setting milestones, and requesting feedback (Nicol & Macfarlane-Dick, 2006). Given this connection between feedback and goals and their critical role in engendering action, we developed indicators of feedback and goal quality and described the extent to which facets of feedback and goals are present or absent.

Methods

Our analyses were driven by the goal of examining the implementation fidelity of the two formative processes of an EES. This provided support for state theory of action that instructional practices are malleable and can be impacted by an EES through principal supervision and feedback. We operationalized this notion by examining written feedback and teacher PGPs to determine, at least preliminarily, the quality of these components. Extant validity evidence (Bell et al., 2012)––particularly g-study evidence of observation rubrics themselves–– exists, but there is little evidence related to feedback based on observations. Although we did not strictly conduct a validation study, Kane's (2013) argument and use approach coupled with Messick's (1995) approach of collecting evidence related to intended inferences guided our thinking in this analysis. Consistent with Messick (1995), we examined the tenability of the state's theory of action using empirical evidence.

We generally followed the procedures outlined in Babbie (2013) to develop the feedback and PGP quality indicators. We first examined the literature to identify facets of good feedback and growth goals, and based on that review, we created items for the indexes. We then used written principal feedback to teachers and teacher written growth plans as artifacts to identify the presence or absence of the various facets of quality. We used a mixed method approach that focused on quantitative analyses and used qualitative methods to illustrate quantitative results (Green & Caracelli, 1989). The qualitative step assisted in evaluating the extent to which feedback based on observations appropriately reflected teachers' strengths and weaknesses (Hill et al., 2012). Based on recommendations in the literature (National Quality Forum, 2013; OECD, 2008; Shwartz & Ash, 2008; Porter, 1991), the quantitative steps included examining internal consistency (Allen &Yen, 1979) and the factor structure (Young & Pierce, 2013). Finally, using the composite scores created for each of the indicators, we examined preliminary relationships using t-tests and correla-

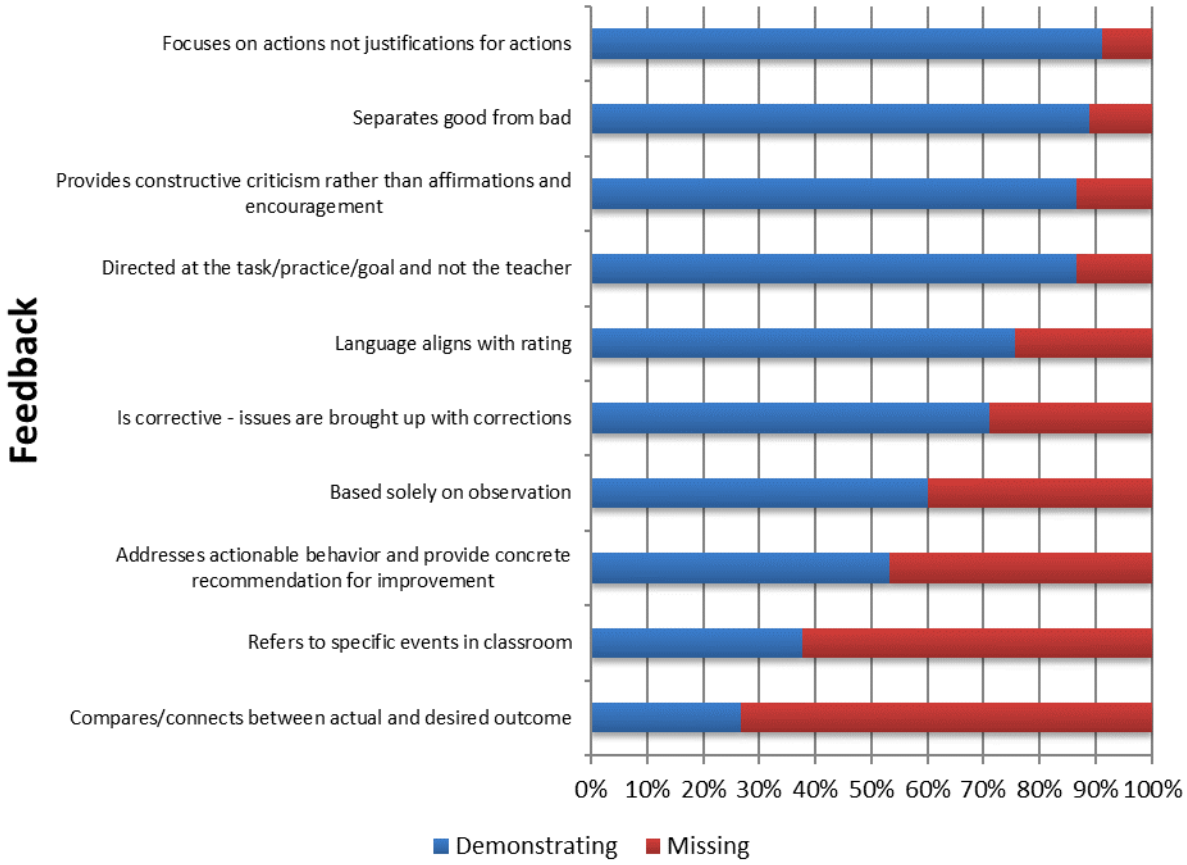

Figure 1. Performance on Feedback Quality Indicator components.

tions. We used a one-way analysis of variance (ANOVA) to test whether there was meaningful between-principal or school variation in the indexes. This analysis provided a limited test that supports previous research that principal impacts vary (Branch, Hanushek, & Rivkin, 2012).

Feedback quality was operationalized by 10 items generally reflecting clarity and utility and qualities that good written feedback ought to exhibit (Heneman & Milanowski, 2004; Ilgen et al., 1979; Kimball & Milanowski, 2009; Kinicki et al., 2004; Meyer, 1991; Smither & Walker, 2004; Thurlings, Vermeulen, Kreijins, Bastiaens, & Stijnen, 2012). The specific items of the FQI are displayed in Figure 1. Although reflections and growth plans are common components of educator effectiveness systems, little research has examined what might constitute a "good" PGP plan. Based on the relevant recent literature (Nicol & MacFarlane-Dick, 2006; Walling, Shapiro, & Ast, 2013) we developed an indicator of PGP quality. The specific items of the PGGI (Professional Growth Goals Indicator) are displayed in Table 1.

Data

We received artifacts consisting of formal observation ratings, written feedback from principals, and PGPs from 42 teachers teaching in an eastern state. The state was in the third year of a statewide EES implementation. The participating districts and schools were recommended by the state education department as being above average implementers of the system[6]. In each school we sought volunteer teachers to provide the requested information, although not all teachers provided all information. Missing data is an important caveat because we had insufficient responses to examine the pattern of missingness or to fully determine how well the sample represented the state as a whole, although observation scores were consistent with statewide results. We were limited by both a small sample and the potential of both non-response and response bias (Babbie, 2013). Given the exploratory nature of these analyses, this is less problematic than it might have been if we attempted to make causal claims. Most teachers (54%) in the sample taught in middle school, were not novice teachers (67%), and did not teach in a STEM field (62%). Despite the limitations, several interesting patterns

Table 1

Professional Growth Goal Indicators and Examples of Those Meeting Criteria

Criteria	Demonstrating
Articulates skill areas to improve upon. "I would love to participate in professional development which will help me improve upon differentiating assignments within a SAM (single approach to mastery) classroom."	41%
Has general PD request. "I would like to attend more content specific professional development workshops to shape my instruction to fit the needs of my students."	28%
Professional goals are clear. "Go back to school and receive a certificate in Educational Technology."	25%
Identifies obstacles. "My ELL period 8 class is my main struggle this year to help them overcome the language barrier and be successful with 9th grade math concepts."	22%
Identifies steps for reaching goal. "My goal this year is to focus on the RTQ Problem solving process. I will be working with my plc teams to analyze student work and identify areas of concern. We will de designing a plan for each student and implementing interventions."	19%
Specifies required actions. "I want to master small group instruction and differentiation. I am attending small group instruction trainings to help me towards this goal."	13%
Provides time line for each action step. "… complete my requirements for my Masters before December. The paper I am writing involves investigating the fairness of school funding in [State]. Help with school funding data would be helpful."	3%
Evaluates current knowledge and skill levels. "I can personally see myself struggling when it comes to instructing reading. I am intimidated by reading due to the fact that kindergarten students have little or no phonemic awareness/phonics skills before entering kindergarten."	3%
Identifies measurable benchmarks.	0%

emerged and warrant continued examination or monitoring.

Results

Table 2 summarizes the overall descriptive results for the indicators we developed as well as the state EES results. Very few teachers had student learning results (and none were used), but most teachers did have observation ratings, which are tightly clustered around the mean of about 3 (satisfactory) with little variation. We address the specific findings in turn, but it is important to note that feedback quality (FQI, FQI2) scores generally averaged about 57% (5.7/10) of possible points and professional growth plans about 13% (1.2/9).

Principal Feedback

We rated principal feedback using the FQI, which consisted of 10 items (Figure 1). Given the prospective nature of the analysis, we scored each item solely on the presence or absence of the construct identified in the item. Additional iterations might develop a scoring rubric that better differentiates feedback on each construct. Overall the FQI appeared to work well. The sample size is insufficient to fully examine the psychometric properties of the indicator; however, internal consistency analysis indicates that the full 10-item FQI has a Cronbach's alpha of .6, while a 9-item version (FQI2) has a reliability of .77.

The difference between FQI and FQI2 is whether feedback made reference to post-observation meeting conversations. This difference reflects variation

Table 2

Mean Indicator and EES Scores

Artifact	*N*	Minimum	Maximum	*M*	*SD*
Student Learning	12	-0.5	1.0	0.5	0.6
Observation Rating	38	2.5	3.6	3.1	0.2
Professional Growth Plan	41	0.0	5.0	1.2	1.2
Feedback Quality (1)	37	0.0	10.0	5.7	3.1
Feedback Quality (2)	37	0.0	9.0	5.2	3.0

among principals in whether written feedback is a summary of observations and discussion or written feedback is a basis for discussion.

Figure 1 presents the distribution of performance on each of the FQI items. The items in Figure 1 are ordered from "easiest" to "hardest;" that is, elements of the FQI that were most readily observed in the feedback are at the top of the figure. For example, we found that most principals were able to provide feedback that focused on actions but not the justification for actions.

Principals had a much more difficult time comparing/connecting actual and desired behavior. The results in Figure 1 clearly indicate that there are attributes to high quality feedback that vary in the propensity of their appearance on written teacher feedback.

We next examined whether the individual responses on the FQI formed relevant and meaningful latent factors, which may be helpful in identifying how principals might benefit from professional development themselves. Although the sample size was small, guidance for conducting exploratory factor analysis varies with recommendations focusing on either an absolute minimum *N* or a subject-to-variable ratio. Minimum *N*'s as low as 40 and subject to variable ratios as low as 2:1 have been utilized in the literature, although this is below the common benchmark of 150 (Young & Pearce, 2013). While there is no specific cut-off, the robustness of results depends to a large extent on the empirical results (Zhao, 2009).

We applied principal component exploratory factor analysis and found that the 10 items presented in Figure 1 behave quite well in forming a two-factor solution[7]. The variance explained for the FQI is consistent with its reliability, about 0.58. The two-factor solution is quite informative and represents two domains: clarity of communication (M = 3.4/4), and instructional practices (M = 3.4/6). The instructional practices factor focuses on specific observed classroom practices, areas for improvement, and specific recommendations for improvement. The clarity of communication factor focuses on communication––the feedback language is aligned to the rating, comments are directed at teachers, and feedback clearly delineates strengths from weaknesses. Results indicate that principals are able to, for the most part, provide feedback that is communicated well, in terms of aligning to the rating and focusing on actions as opposed to the person. However, principals seem less able to consistently provide specific feedback with concrete examples from the classroom that are linked to areas of improvement and are aligned to desired outcomes and specific recommendations as to how to achieve the desired outcomes.

In order to solidify the concepts presented above, we provide specific examples from the feedback forms. Although principals generally did well in communication, and despite language aligning with ratings about 75% of the time, there are several examples of misalignment. One principal stated that the teacher "did not get to cover what she wanted in the lesson because time ran short. It has been recommended that she use a visual timer." However, this teacher scored *distinguished* in maximizing learning time on the evaluation rubric.

In another example of misalignment, a principal communicated,

> It is recommended that you use formative assessment to gauge student progress . . . you did not directly assess understanding of the text prior . . . additionally, you did not present a summarizing task . . . recommended that you devote the majority of your instructional time to content-related learning tasks.

Despite this feedback, this teacher scored *proficient* across all elements in the evaluation rubric.

In terms of feedback related to instructional practices, principals had a significantly more difficult time providing concrete guidance. For example, in terms of constructive criticism rather than affirmations and encouragement, one principal stated, "It is recommended that you continue this program with fidelity." This provides no constructive criticism and focuses on affirmation.

Some principals did provide a straightforward example of meeting this criterion, as one asserted, "In order to move to a distinguished level have students plan to ask 1-2 questions after they compared their markings."

Although occurring less than 40% of the time, some feedback did refer to specific events in classroom, such as when a principal stated,

> As we discussed in the post-conference, you not only gave students recall questions to answer as they read, but you told them exactly where to find the answers. When giving students an important text to read, determine your purpose first and then provide an appropriate graphic organizer and/or require the use of an effective reading strategy that promotes deeper understanding of the text.

Principals also did present concrete issues with corrective actions, as in a principal's feedback:

> Prepare to move the lesson along when/if students are able to grasp concepts more quickly than anticipated. Students appeared to quickly understand the significance of a PSA and the components of an effective PSA. More time can be spent on student production of their PSA related to toxins.

While occurring about a quarter of the time in written feedback, comparison/connection between actual and desired outcome was present in some feedback. For example, a principal suggested, "Have the students to share their data for finding right angles instead of her sharing that information. That would have given students who did not finish the activity [opportunity] to complete the task as well."

We examined both characteristics of the teachers and of the observation to determine whether there were any systematic relationships with FQI scores. Overall, the average teacher rating on the observation protocol (FFT) was inversely related to the FQI (r = -.27, p < .10)[8], despite the lack of variability in the FFT[9]. Principal feedback was not qualitatively different whether the observation was announced (40% of the observations) or unannounced. Overall, feedback quality did not differ between novice and experienced teachers; however, feedback related to instructional practice (the instructional practices factor) was of significantly lower quality for novice teachers (d = .15, p < .05).

Importantly there is evidence that the number of elements of the FFT scored relates to both teacher overall ratings[10] and the quality of feedback they receive. Teachers scored on fewer elements of the FFT tended to have higher overall ratings (r = -.32, p < .05). The overall FQI is positively related to the number of elements scored (r = .51, p < .01). Each domain of the FQI is related to the number of elements scored. The instructional practices factor is positively related to the number of elements scored (r = .50, p < .01), and the clarity of communication factor is positively related to the number of elements scored (r = .49, p < .01).

Table 3 summarizes the variability of the quality of feedback. In other words, Table 3 provides some evidence as to whether there are statistically significant differences in feedback among evaluators and schools[11]. There is suggestive evidence that the quality of feedback, particularly instructional feedback quality, varies by evaluator. The results in Table 3 suggest that there tends to be systematic differences among schools in teacher ratings as well as the quality of the feedback. Differences among schools represent either mean differences in teachers, rater stringency, or feedback quality[12].

Table 3
Variation in Principal Feedback

	Evaluator	School
Overall Teacher Rating	no	p < .10
Instructional Prac.	p < .10	p < .05
Clarity of Comm.	no	No
FQI	no	p < .01
FQI2	p < .10	p < .01

Note. As variation assessed by a one-way ANOVA with evaluators or schools as the groups.

Professional Growth Goals

We next examined teachers' Professional Growth Goals (PGGs). We note that this form was not being completed or evaluated with fidelity. There was no guidance provided to teachers, and there was no place for evaluator comments on the form. We evaluated teachers' written plans using a set of items (Professional Growth Goal Indicator [PGGI]) derived from the literature considered to meaningfully

describe aspects of quality related to goal setting. Each teacher's goal was scored on the nine items presented in Table 1. Given the limited sample size and the exploratory nature of this construct, we simply coded for the presence or absence of the element. In this way, we were able to determine the extent to which teachers, without guidance, were able to develop a quality growth plan.

The results in Table 1 clearly indicate that teachers' reflections only loosely develop professional plans. Less than 50% of the plans identified a specific skill area to improve upon. In terms of specific actions required to meet goals, less than 20% of plans specified actions, and no plan identified a measurable benchmark that would provide evidence that the goal had been met. These results provide evidence as much for the need to provide concrete direction as they do for teachers' inability to develop coherent growth goals.

Overall, the average PGG scored about a 1.2 out of a possible 9 points. These low scores impact the reliability of the instrument because the modal score was 0. Additional research is required to determine whether the instrument is incapable of identifying the distribution of quality in growth plans, or whether, in fact, growth plans are not developed with fidelity across the state.

While unequivocal claims about plans would be unjustified, substantive evidence does indicate that the plans are not completed with fidelity. For example, one plan's articulated goal was to "be the best teacher I can," while another plan indicated that the goals were to "continue to learn as a teacher . . . and attend workshops."

On the other hand, there were examples of plans meeting specific criteria. Table 3 presents representative samples from growth plans and the proportion of plans that met the criteria. In some instances, there were few exemplars from which to choose. In addition, providing measurable benchmarks was not indicated on any plan consistently[13].

Despite the limited range in overall teacher ratings and scores on the PGGI, there is evidence that more effective teachers wrote stronger growth goals, shown by a positive correlation between overall teacher ratings and growth goal scores ($r = .37$, $p < .05$). Likely consistent with expectations is that principal feedback was inversely related to growth goal scores ($r = -.35$, $p < .05$). This indicates that teachers who wrote better plans (who tend to be more highly rated teachers) received lower quality feedback.

It is also interesting to note that the strength of goals did not vary systematically among the schools in the sample, but that two schools in the sample had means that were two to three times higher than the other two schools in the sample. The difference between the two pairs of schools was significant ($p < .05$). Again we note that these results are not based on simple random sample and that inferences based on statistical tests should be considered with caution.

Although there is insufficient evidence to support the notion that growth goals vary systematically among schools, in general, there are suggestive patterns to the results. For example, while the majority of plans at *School A* articulated a skill area to improve upon, all *School A* plans missed the same six criteria. A similar (yet different) pattern existed for *School B*. Additional investigation can examine whether teachers in these schools were provided specific direction or guidance (that coincidentally met some of the criteria applied in this evaluation). The other two schools seemed to demonstrate more variability among the plans––with less concentration on particular aspects, but broader coverage. Together the results suggest that there may be differences among the schools in how they approached growth plans and that developing guidance and policy can impact how teachers address this task. While there was a positive relationship between overall teacher rating and the quality of plans, there were no significant correlations within schools. The point estimates of the correlations varied from .23 to .45, but there was insufficient sample size within schools to detect relationships. Again, the results provide suggestive evidence for school-wide differences in approaching PGGs.

Discussion

There has been a recent emphasis on transformational leadership skills of principals (Bluestein, 2011), and current EESs renew a focus on supervision and evaluation to facilitate improvement in instructional quality. Recent research indicates that EESs can impact instruction quality, as measured by student outcomes.

However, to engender the large-scale effects that states and districts are hoping for, it is important to consider that feedback works best when decoupled from evaluation (Marshall, 2013; Meyer, 1991), and this is confirmed by teachers who indicate that this is precisely what they prefer (Marshall, 2013). It is also important to consider that teachers rate the feedback they receive from principals less useful than principals rate the feedback they provide (Hallinger & Heck, 1996). Fortunately, principals desire additional men-

toring and professional development specifically on providing feedback (Johnston, J. Kaufman, & L. Thompson, 2016).

Our initial examination of feedback indicates that feedback tends to fall along two dimensions. One dimension focuses on clarity of communication. This dimension of feedback provides results related to how well principals' feedback is written––whether it is clear and objective. This is an important dimension since clearly communicated feedback focusing on tasks rather than traits enhances recipients' perception of accuracy (Kinicki et al., 2004; Kluger & DeNisi, 1996); principals generally do a good job in this dimension. The second dimension emphasizes instructional practices. This dimension focuses on feedback using specific classroom practices to highlight strengths and weaknesses and to develop concrete recommendations for improvement as well as strategies to engender that improvement. This dimension is important because it is through focusing attention that feedback engenders change (Larson et al., 2013); principals were less successful at providing this sort of feedback. Both more effective teachers and novice teachers received lower quality instructional feedback. Our evidence suggests that the fewer FFT elements scored, the poorer the quality of the feedback. This suggests that a series of mini-observations (Marshall, 2013) may facilitate principals' abilities to observe a broader array of FFT elements over the course of time, increasing the potential for effective feedback. Overall, these results suggest that districts should provide additional mentoring and support specific to instructional practices. This might include helping the principal develop additional knowledge about strategies to address specific instructional issues or creating a clearinghouse or repository to which principals have ready access.

More concrete feedback, which is included in domain four of the FFT (Danielson, 1996), specifically to provide guidance to teachers as well as to provide support in developing growth goals, could lead to more concrete actionable and monitorable goals that can meaningfully guide teachers to professional improvement. Our results indicate that, based on the criteria applied, teachers generally did not develop succinct growth goals that incorporated concrete steps as well as measureable benchmarks for success.

Limitations and Recommendations

This exploration into the fidelity of implementing the formative components of an EES is limited by several key factors. One is the sampling plan, which relied on volunteer responses. Although the results suggest that teacher observation scores were in-line with state performance, it is unknown what unobserved factor related to feedback is related to teachers volunteering their feedback for study. Additionally, the sample size was relatively small given the number of analyses we conducted. As noted, we view the results provisionally, but we also believe that they are consistent with anecdotal evidence from the field. A final limitation is the lack of raw student performance data. We had access to principals' ratings of teachers (based on various student learning results and algorithms), but these ratings included a subjective element, were attenuated, and lacked variability.

The preliminary results do highlight areas for consideration by states. If, in fact, feedback is generally lacking in specifics for improving instructional practices, then this clearly limits the ability of an EES to impact student outcomes. Professional development for principals would be warranted. This is consistent with previous research suggesting that guidance for providing feedback would be beneficial (Johnston, Kaufman, & Thompson, 2016; Scheeler et al., 2004). Importantly, teachers need access to resources and training (Kimball, 2002), and feedback is only helpful if teachers have enabling conditions (DuFour & Marzano, 2009; McLaughlin & Pfeifer, 1988. The variability in quality by principals and schools warrants additional research to determine which principal and district factors contribute to this variation. The lack of fidelity with which growth goals were completed clearly points to a need for additional thinking in providing opportunities and rationale to solidify growth goals as concrete actionable plans with measurable objectives that teachers meaningfully ascribe to and use to monitor their own success towards continued improvement. The systematic variation in PGPs by schools provides some evidence that guidance varied systematically, and it implies that teachers responded to that guidance.

Moreover, if EESs explicitly followed feedback and growth goals as a basis for subsequent evaluation cycles and if principals were also monitored on feedback quality, this may enhance the fidelity with which each of these elements is implemented, furthering the potential to improve instructional practice.

Notes

[1]Feedback need not necessarily come from the principal, but given its prevalence we simply use principal to refer to principals or any other evaluator that provides feedback.

[2]We use the terms Professional Growth Plan and Professional Growth Goals interchangeably.

[3]Links refer to a representative sample of five states: http://www.mde.k12.ms.us/docs/teacher-center/teacher-evaluation-modifications-for-2014-2015.pdf?sfvrsn=2; http://www.nj.gov/education/AchieveNJ/teacher/approvedlist.pdf; http://www.connecticutseed.org/wp-content/uploads/2014/10/CCT_Rubric_for_Effective_Service_Delivery_2014.pdf; http://www.nctq.org/docs/NC_teacher-eval_process.pdf; and http://www.ped.state.nm.us/ped/NMTeach_EvaluationPlan.html.

[4]For example, Self-Determination Theory, (Deci and Ryan, 2000) and Control Theory (Taylor, Fisher, & Ilgen, 1984) have been posited to relate to feedback response and goal setting.

[5]Like the majority of studies on feedback in education, Sadler focuses on teachers and students.

[6]This introduces a limitation on generalizability. It should be noted however, that the state's definition of implementation was related to compliance not to fidelity (to which this study provided some evidence).

[7]The empirical evidence suggests that the results appear well behaved (Young and Pearce, 2013), particularly with no cross-loaded factors and each of the two factors having at least 2 items with loadings over 0.8, and the remaining loadings of around 0.7.

[8]This relationship is significant when using FQI2 (r = -.32, p < .05).

[9]There is little variation in observation ratings in the sample: Observations are distributed with M = 3.05 (effective) and SD = .17.

[10]We use the average score on the 4-point rubric and not the classifications used on the EES summative reports.

[11]Results are based on 11 evaluators and 4 schools.

[12]There is insufficient data to fully examine the structure of the relationships (e.g., the between rater, within school variability).

[13]In a few cases teachers indicated that they intended to get another degree. We did not score these as having benchmark given the lack of specificity in presenting this goal. This is an instance where a more sophisticated scoring rubric would be beneficial (e.g., no benchmark = 0, loose = 1, and concrete = 2).

References

Allen, M., & Yen, W. (1979). *Introduction to measurement theory*. Boston: Brooks/Cole.

Babbie, E. (2013). *The practice of social science research*. Australia: Wadsworth.

Bell, C., Gitomer, D., McCaffrey, D, Hamre, B., Pianta, R., & Qi, Y. (2012). An argument approach to observation protocol validity. *Education Assessment, 117*, 62-87.

Bjorn, K., Wurth, S., & Hergovich, A. (2013). The impact of feedback on goal setting and task performance. *Swiss Journal of Psychology*, *72*(2), 79-89.

Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education: Principals, Policy, and Practice*, *5*(1), 7-68.

Bluestein, S. (2011). *Principal effectiveness in California elementary schools* (Doctoral dissertation). Retrieved from CSUN Electronic Theses and Dissertations.

Branch, G., E. Hanushek, & Rivkin, S. (2012). *Estimating the effect of leaders on the public sector productivity: The case of school principals.* (Working Paper No. 66). Washington, DC: American Institutes for Research.

Chetty, R., Friedman, J., & Rockoff, J. (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *America Economic Review*, *104*(9), 2593-2632.

Chetty, R., Friedman, J., & Rockoff, J. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *America Economic Review*, *104*(9), 2563-2679.

Cianci, A. M., Klein, H. J., & Seijts, G. H. (2010). The effect of negative feedback on tension and subsequent performance: The main and interactive effects of goal content and conscientiousness. *Journal of Applied Psychology, 95*(4), 618-630.

Danielson, C., & McGreal, T. (2000). *Teacher evaluation to enhance professional practice*. Alexandria, VA: Association for Supervision and Curriculum development.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum development.

Darling-Hammond, L. (2004). Inequality and the right to learn: Access to qualified teachers in California's public schools. *Teachers College Record, 106* (10), 1936-1966.

DuFour, R., & Marzano, R. (2009). High-leverage strategies for principal leadership. *Educational Leadership, 66*(5), 62-68.

Eisner, E. (1992). Education reform and the ecology of schooling. *Teachers College Record, 93*, 610-627.

Greene J., Valerie, J., & Caracelli, G. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, *11*, 255–274.

Hahnel, C., & Jackson, O. (2012). *Learning denied: The case for equitable access to effective teaching in California's largest school district.* Oakland, CA: The Education Trust.

Hallinger, P., & Heck, R. H. (1996). Reassessing the principal's role in school effectiveness: A review of empirical research, 1980-1995. *Educational Administration Quarterly*, *32*, 5-44.

Hallinger, P., & Murphy, J. (1985). Assessing the instructional management behavior of principals. *The Elementary School Journal*, *66*(2), 217-247.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81-112.

Heneman III, H., & Milanowski, A. (2004). Alignment of human resource practices and teacher performance competency. *Peabody Journal*, *79*(4), 108-125.

Hill, H., Charalambous, C., Blazar, D., McGinn, D., Kraft, M., Beisiegel, M., Hunnez, A., Litke, E., & Lynch, K. (2012). *Educational Assessment*, *17*, 83-106.

Hysong, S., Best, R., & Pugh, J. (2006). Audit and feedback and clinical practice guideline adherence: making feedback actionable. *Implementation Science, 1*,9.

Hysong, S. (2009). Meta-analysis: Audit and feedback features impact effectiveness on care quality. *MedCare*, *47*(3), 356-63.

Ilgen, D., Fisher, C., & Taylor, S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology, 64*(4), 349-371.

Johnston, W., Kaufman, J., & Thompson, L. (2016). *Support for instructional leadership: Supervision, mentoring and professional development for US school leaders: Findings from the American School Leader Panel.* Santa Monica, CA: RAND Corporation.

Kane, M. (2013). Validating the Interpretation and Use of Test Scores. *Journal of Educational Measurement*, *50*(1), 1-73.

Kane, T., McCaffrey, D., Miller, T., & Staiger, D. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T., & Staiger, D. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Practioner_Brief.pdf

Kane, T., Taylor, E., Tyler, J., & Wooten, A. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources*, *46*(3), 587-613.

Kimball, S., & Milanowski, A. (2009). Examining teacher evaluation validity and the leadership decision making within a standards-based evaluation system. *Educational Administrative Quarterly*, *45*(1), 34-70.

Kimball, S. (2002). Analysis of feedback, enabling conditions and fairness perceptions of teachers in three school districts with new standards-based evaluation systems. *Journal of Personnel Evaluation in Education, 16*(4), 241-268.

Kinicki, A., Wu, B., Prussia, G., & McKee-Ryan, F. (2004). A covariance structure analysis of employees' response to performance feedback. *Journal of Applied Psychology*, *89*(6), 1057-1069.

Kluger, A., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254-284.

Larson, E., Patel, S., Evans, D., & Saiman, L. (2013). Feedback as a strategy to change behavior: The devil is in the details. *Journal of Evaluation in Clinical Practice, 19*(2), 230–234.

McLaughlin, M. & Pfeifer, R. (1988). *Teacher evaluation: Improvement, accountability, and effective learning.* NY, NY: Teachers College Press.

Marshall, K. (2013). *Rethinking teacher supervision and evaluation* (2nd ed.). San Francisco: Jossey-Bass.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Meyer, H. (1991). A solution to the performance appraisal feedback enigma. *Academy of Management Executive*, *5*(1), 68-76.

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education*, *79*(4), 33-53.

Morey, E. (2003). Feedback research revisited. In D. Jonasson (ed.), *Handbook of research for educational communication and technology* (pp.745-784). New York: McMillan.

Nicol, D., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Learning, 31*(2), 199-218.

National Quality Forum (2013). *Composite performance measure evaluation guidance.* Washington, DC: National Quality Forum.

Nye, B., Konstantopoulos, S., & Hedges, L. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*(3): 237-257.

OECD. (2008). *Handbook on Constructing Composite Indicators: Methodology and Users Guide*. Paris: OECD.

Porter, A. (1991). Creating a system of school process indicators. *Educational Evaluation and Policy Analysis*, *13*(1), 13-29.

Rivkin, S., Hanushek, E., & Kain. J. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417-458.

Rose, D., & Farrell, T. (2002). *The use and abuse of comments in 360-degree feedback.* Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Toronto, Canada.

Sadler, R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18* (2), 119-144.

Scheeler, W., Dochy, F., & Janssens, S. (2004). Providing performance feedback to teachers: A review. *Teacher Education and Special Education*, *27*(4), 59-70.

Shepard, L. (2012). *Evaluating the use of tests to measure teacher effectiveness: Validity as a theory-of-action framework*. Paper presented at the National Council of Measurement in Education Meeting, Vancouver, Canada.

Shwartz, M. & Ash, A. (2008). *Composite measures: Matching the method to the purpose.* Agency for Healthcare Research and Quality. Retrieved from http://www.qualitymeasures.ahrq.gov/expert/expert- commentary.aspx?id=16464

Smither, J., & Walker, A. (2004). Are the characteristics of comments related to improvement in multirater feedback ratings over time?. *Journal of Applied Psychology, 89*(3), 575-581.

Steinberg, M., & Donaldson, M. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy*, *11*(3), 340–359.

Stiggins, R., & Duke, D. (1988). *The case for commitment to teacher growth: Research on teacher evaluation.* Albany, NY: State University of New York Press.

Taylor, E., & Tyler, J. (2012). The effect of evaluation on teacher performance. *American Economic Review*, *102*(7), 3628-3651.

Thurlings, M, Vermeulen, K., Kreijins, K., Bastiaens, T., & Stijnen, S. (2012). Development of the teacher feedback observation scheme: Evaluating the quality of feedback in peer groups. *Journal of Education for Teaching: International Research and Pedagogy*, *38*(2), 193-208.

Walling, A., Shapiro, J., & Ast, T. (2013). What makes a good reflective paper?. *Family Medicine*, *34*(1), 7-12.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness.* New York, NY: The New Teacher Project.

White, S. (2009). Articulation and re-articulation: Development of a model for providing quality feedback to pre-service teachers on practicum. *Journal of Education for Teaching*, *35*(2), 123-132.

Young, A., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory analysis. *Tutorials in Quantitative Methods for Psychology, 9* (2), 79-94.

Zhao, N. (2009). *The minimum sample size in factor analysis.* Retrieved from https://www.encorewiki.org/display/~nzhao/The+Minimum+Sample+Size+in+Factor+Analysis