Organizational Behavior and Human Decision Processes
Vol. 85, No. 1, May, pp. 32–55, 2001
doi:10.1006/obhd.2000.2931, available online at http://www.idealibrary.com on IDEAL®

The Assessment of Goal Commitment: A Measurement Model Meta-Analysis

Howard J. Klein

The Ohio State University

Michael J. Wesson

Texas A&M University

John R. Hollenbeck

Michigan State University

Patrick M. Wright

Cornell University

and

Richard P. DeShon

Michigan State University

Goals are central to current treatments of work motivation, and goal commitment is a critical construct in understanding the relationship between goals and performance. Inconsistency in the measurement of goal commitment hindered early research in this area but the nine-item, self-report scale developed by Hollenbeck, Williams, and Klein (1989b), and derivatives of that scale, have become the most commonly used measures of goal commit-

The authors thank Jennifer Kantrowitz and Brenda LeVan for their assistance in compiling the data used in this study and Jason A. Colquitt and the anonymous reviewers for their valuable comments on earlier versions of this paper. A version of this paper was presented at the Annual Meeting of the Academy of Management. Chicago, Illinois, August, 1999.

Address correspondence and reprint requests to Howard J. Klein, Department of Management and Human Resources, Max M. Fisher College of Business, The Ohio State University, 2100 Neil Avenue, Columbus, OH 43210–1144. E-mail: klein.12@osu.edu.

0749-5978/01 $35.00

ment. Despite this convergence, a few authors, based on small sample studies, have raised questions about the dimensionality of this measure. To address the conflicting recommendations in the literature regarding what items to use in assessing goal commitment, the current study combines the results of 17 independent samples and 2918 subjects to provide a more conclusive assessment by combining meta-analytic and multisample confirmatory factor analytic techniques. This effort reflects the first combined use of these techniques to test a measurement model and allowed for the creation of a database substantially larger than that of previously factor analyzed samples containing these scale items. By mitigating sampling error, the results clarified a number of debated issues that have arisen out of previous small sample factor analyses and revealed a five-item scale that is unidimensional and equivalent across measurement timing, goal origin, and task complexity. It is recommended that this five-item scale be used in future research assessing goal commitment.

Goals have emerged as a central, pervasive construct in the motivation literature (Austin & Vancouver, 1996) and task goal theory (Locke & Latham, 1990) has received the bulk of this attention. Within goal theory, goal commitment has been identified as an essential condition since a goal can have no motivational effect if there is no commitment (Locke, Latham, & Erez, 1988). Measurement inconsistency was an issue raised in early reviews of the goal commitment literature (Hollenbeck & Klein, 1987; Locke et al., 1988). The use of single item measures has since largely given way to the use of the self-report scale developed by Hollenbeck, Williams, and Klein (1989) (referred to henceforth as the HWK scale). A recent meta-analysis of goal commitment research concluded that, while goal commitment has begun to receive greater attention, considerably more research on goal commitment is warranted given its critical role in goal theory (Klein, Wesson, Hollenbeck, & Alge, 1999). With this need for additional research, it is imperative that the measure used to assess goal commitment be construct valid and demonstrate appropriate psychometric properties.

Despite the widespread use of the HWK measure to assess goal commitment, a few articles have questioned the dimensionality of this measure (e.g., DeShon & Landis, 1997; Tubbs, 1993). To address the conflicting recommendations in the literature and the resulting confusion that may exist regarding which HWK items to use when measuring goal commitment, a collaborative effort was undertaken to test the adequacy of this measure using recent methodological advances, namely, the combination of meta-analytic techniques and structural equation analyses (Viswesvaran & Ones, 1995). This approach allows a robust assessment of the most commonly used measure of goal commitment and the clarification of a number of issues that have arisen out of previous small sample studies. The purpose of this examination is to identify a unidimensional measure of goal commitment that researchers can confidently use in the future.

CONSTRUCT DEFINITION

Goal commitment is defined as *"one's determination to reach a goal"* (Locke & Latham, 1990). This broad definition of goal commitment is consistent with the current conceptualization of the construct within task goal theory. Implicit in this definition is the intention to extend effort toward goal attainment, persistence in pursuing that goal over time, and an unwillingness to lower or abandon that goal (Hollenbeck & Klein, 1987). While there is general consensus within goal theory on this definition, it should be noted that there are exceptions. Tubbs and Ekeberg (1991) drew finer conceptual distinctions between initial agreement with a goal as a standard (i.e., acceptance), intentions to attain a goal (i.e., commitment), and the maintenance of those intentions over time (i.e, determination). While these subtle conceptual distinctions can be made, all three are motivational choices thought to be influenced by the same antecedents and previous research has failed to show such distinctions to be empirically useful (Locke & Latham, 1990).

One of the major conclusions from the Hollenbeck and Klein (1987) review was that goal commitment rarely played a major role in empirical studies despite being a central construct within goal setting theory. Goal commitment was typically left unmeasured or was measured with a single item as a means for eliminating subjects who reported low commitment. The HWK scale, shown in Table 1, was developed by first examining how goal commitment had been defined and measured in the goal setting literature. As noted above, a variety of single item measures were in use at the time. Some of those previously used single item measures, judged to be consistent with the construct as defined above, were incorporated into the HWK scale. Additional items, again consistent with the conceptual definition of the construct, were written to round out the scale. The origin of each item is described further in Hollenbeck, Klein, O'Leary, and Wright (1989a).

TABLE 1

Hollenbeck, Williams, and Klein (1989b) Goal Commitment Items

1. *It's hard to take this goal seriously.* (R)
2. It's unrealistic for me to expect to reach this goal. (R)
3. It is quite likely that this goal may need to be revised, depending on how things go. (R)
4. *Quite frankly, I don't care if I achieve this goal or not.* (R)
5. *I am strongly committed to pursuing this goal.*
6. *It wouldn't take much to make me abandon this goal.* (R)
7. *I think this is a good goal to shoot for.*
8. I am willing to put forth a great deal of effort beyond what I'd normally do to achieve this goal.
9. There is not much to be gained by trying to achieve this goal.

Note. Items followed by "R" indicate that the item was reverse-scored before analysis. Italicized items are those included in the recommended five-item scale.

DIMENSIONALITY OF THE HWK SCALE

Initial construct validity evidence for this measure was provided by Hollenbeck et al. (1989a). Based on a sample of over 750 subjects, they examined the dimensionality and reliability of the scale along with the convergence and divergence from other theoretically relevant and irrelevant constructs. The results of a principal components factor analysis for the full nine-item HWK scale revealed one major underlying factor, although a second factor, accounting for a small amount of additional variance, was also evident. Shorter, seven- and four-item versions of the scale were also examined and both of the shorter versions were unidimensional. Hollenbeck, Klein, et al. (1989a) thus recommended the use of the four-item scale as a manipulation check and the use of the seven-item scale for research examining goal commitment as a construct of interest.

Despite the convergence in the literature on the use of this scale and the 10 years of accumulated empirical evidence that the scale relates to other constructs in a manner consistent with its conceptual definition (Klein et al., 1999), concerns have been raised about the dimensionality of the HWK measure (e.g., DeShon & Landis, 1997; Tubbs, 1993). The issue of whether the items in the HWK scale might tap multiple conceptual dimensions was first raised by Tubbs and his colleagues (Tubbs, 1993; Tubbs & Dahl, 1991; Tubbs & Ekeberg, 1991). Taking the position that the conventional definition of goal commitment is too broad, Tubbs and Dahl (1991) subjectively categorized the HWK scale items, along with other items that have been used, into more narrow conceptual definitions. Empirically, when they factor analyzed the nine-item HWK scale, they found evidence for a two-factor solution. Those two factors did not, however, reflect the conceptual distinctions they had made.

DeShon and Landis (1997) suggested that only a subset of the items in the HWK scale were consistent with the broader, more conventional, definition of goal commitment. DeShon and Landis (1997) presented results from three empirical studies, two using eight of the nine HWK items, the third using the seven-item scale. The findings from all three studies suggested a two-factor solution, with the second factor consisting of three items (items 1–3) which they interpreted as representing the cognitive evaluation of goal attainment (i.e., expectancy evaluations) rather than goal commitment. Given that the literature currently has different groups of authors making different assertions regarding the dimensionality of the HWK scale, researchers may be understandably confused over which items, if any, to use when measuring goal commitment.

It should be noted that two separate camps exist within our author team. Rather than go the typical rebuttal route, we chose to work together to identify a unidimensional set of items that researchers can confidently use in the future. We were also in agreement that the sample sizes used in the Tubbs and Dahl (1991) study (n = 125) and all three DeShon and Landis (1997) studies (n = 85, 88, and 100) were smaller than what is desirable for confirmatory factor

analysis. As such, this study combines meta-analytic techniques and multisample confirmatory factor analyses to reexamine the factor structure of the full nine-item HWK scale and test the adequacy of the HWK measurement model. Doing so will determine whether the scale is multidimensional or whether the occasional failure to replicate a unidimensional structure is a function of sampling error.

While we agreed that a consensus conclusion on an appropriate, unidimensional scale could be reached through this approach, we did not attempt to reach an *a priori* consensus regarding which of the nine items would be in that scale. Those associated with the original HWK scale felt that all nine items conceptually reflect the definition of goal commitment but, given the accumulated empirical evidence, the seven-item scale (items 1–7) should be used. The DeShon and Landis position was that a five-item subset (items 4–8) should be used. Given this overlap in prior recommendations, there was implicit agreement on the inclusion of items 4–7 and the exclusion of item 9. The disagreement centered primarily on items 1–3. These items are part of the seven-item HWK scale but DeShon and Landis (1997) argued that these items assess a separate construct. Therefore, *the primary purpose of this study was to derive a unidimensional, robust measure of goal commitment from the nine HWK scale items.*

SCALE GENERALIZABILITY

In trying to explain why the HWK scale has occasionally been found to reflect two dimensions, DeShon and Landis (1997) hypothesized task complexity as a potential moderator. While they did not examine tasks of varying complexity, they suggested that the simplistic nature of tasks often used in goal setting research may have contributed to the unidimensionality commonly observed with the HWK scale. Recall that DeShon and Landis (1997) suggested that their second factor reflected expectancy evaluations. They argued that on relatively simple tasks, performance is largely a function of effort and, as such, the relationship between expectancies and commitment should be quite high. In contrast, the association between effort and performance is weaker for complex tasks and the relationship between expectancies and commitment should also be weaker. Given that rationale, if the HWK scale derived in addressing the primary purpose of this study is multidimensional, such that some of the items reflect expectancy evaluations rather than goal commitment, distinct patterns for those items should be evident for complex tasks but not for simpler tasks. Alternatively, if the derived HWK scale is unidimensional, the structure of the scale should be equivalent across tasks of varying complexity. Therefore, the following research question was examined: *Is the structure of the derived scale equivalent across tasks of varying complexity?*

While it is important for a measure of goal commitment to be valid across a variety of tasks, it is also important for the measure to be robust across different temporal periods. Because commitment implies persistence in pursing a goal over time, the timing of the measure should not affect the dimensionality

of the scale. While the level of commitment reported by a respondent can be expected to change over time, the structure of the commitment measure must remain stable so that changes in commitment over time can be assessed. Researchers have used the HWK scale to assess goal commitment before, during, and even after goal striving. Significant effects for the timing of measurement have not been evident in past research looking at relationships between goal commitment and other constructs (e.g., Earley & Kanfer, 1985; Hollenbeck, Klein et al., 1989a). Previous research has not, however, examined the potential effects of timing on the structure of the scale itself. Tubbs (Tubbs & Ekeberg, 1991; Tubbs, 1993) suggested that initial intentions and the maintenance of those intentions are distinct constructs requiring separate measures. If the scale is actually multidimensional, with different dimensions assessing different temporal periods, then differential factor structures should be evident in studies assessing commitment prior to goal striving versus studies assessing commitment during goal striving. Alternatively, if the derived HWK scale is unidimensional, the structure of the scale should be equivalent across the timing of the measure. Therefore, the following second research question was examined: *Is the structure of the derived scale equivalent across the timing of the measurement?*

A third aspect of the context that might influence the dimensionality of the scale is goal origin. Hollenbeck, Williams et al. (1989b) and other researchers have used the HWK scale to assess commitment to both assigned and self-set goals. As with the timing issue, Tubbs (1993) has suggested that one's adherence to an assigned goal and the strength of one's personal intentions are distinct constructs requiring separate measures. If the HWK scale is multidimensional, assessing separate acceptance and commitment constructs, as suggested by Tubbs and Dahl (1991), then different factor structures should be evident in studies using self-set versus assigned goals. Alternatively, if the derived HWK scale is unidimensional, the structure of the scale should be equivalent across goal origin. The following final research question was thus examined: *Is the structure of the derived scale equivalent across goal origin?*

ADDITIONAL ISSUES

While the structure of the derived scale should be stable across settings and contexts, differences in scale level or variation could be expected and would not invalidate the scale. Such differences will be examined to provide additional validity evidence for the scale (to the extent to that observed differences are consistent with goal theory) and to provide more detailed normative information (providing future researchers with information on scale properties under a variety of contexts). Finally, to complement the empirical tests of the measurement model and to confirm the subjective judgments of the authors and the those of the naive subjects who participated in the included studies, a panel of expert judges was used to provided additional evidence regarding the conceptual correspondence between the goal commitment construct and the HWK items to ensure that the resulting scale is still construct relevant.

METHOD

Analytical Approach

The current study is unique in combining meta-analytic techniques and multisample confirmatory factor analyses to test the adequacy of a measurement model. Meta-analysis was used to combine multiple small sample studies in order to mitigate the effects of sampling error. Hunter and Schmidt (1990) have demonstrated how inferences drawn from small sample studies can lead to highly erroneous conclusions when taken in isolation. While meta-analysis is typically used to aggregate effect sizes between two constructs, sampling error is also a potential alternative explanation for discrepant findings regarding interitem effects for a particular measure. Based upon the results from several small sample studies, some have concluded that the evidence regarding the HWK scale is conflicting and equivocal. Meta-analysis greatly reduces the potential of sampling error as an alternative explanation for the discrepant findings in those small sample studies. An aggregate interitem covariance matrix was computed from the results of a series of item-level meta-analyses. That covariance matrix was then used as the input for factor analyses examining the dimensionality of the HWK scale. Multiple groups confirmatory factor analysis was then used to test the equivalence of the scale across samples differing in goal origin, measurement timing, and task complexity. Each of these analyses is explained in greater detail in the paragraphs that follow.

Studies Used in the Analysis

Most meta-analyses begin with a literature search for two or more constructs so that a cumulative assessment can be made regarding the relationships between those constructs. The present analysis was substantially more focused, involving only a single construct and a specific measure of that construct. Specifically, the following inclusion criteria were used for the current meta-analyses: (a) commitment to an individual-level goal had to be measured using at least seven of the nine items that make up the HWK self-report measure of goal commitment and (b) suitable statistics had to be available in order to calculate the mean, standard deviation, and interitem correlations for each of the items.

A common problem with meta-analysis is that published studies often do not report the statistics needed to determine the effect size between the variables of interest (Hunter & Schmidt, 1990). This problem is exacerbated in the current analysis as *item-level statistics* are required. Because most published studies do not report the interitem correlations for the measures they use, a representative sample of research using the HWK scale could not be obtained using traditional meta-analysis search procedures. In fact, a traditional search using both computer and manual methods resulted in the identification of only one article reporting the necessary statistics. Even in that article (DeShon & Landis,1997), scale interitem correlations were provided for only one of the three studies reported.

Since published articles do not report the necessary information, an alternative approach was used to identify a representative sample of data sets meeting the above stated inclusion criteria. Three different researchers who had published studies using the HWK measure were asked to provide select information from any available data sets meeting the inclusion criteria described above. In total, this procedure produced data from 17 independent samples representing a variety of settings, respondents, and conditions. Sample sizes ranged from 59 to 494 ($M = 171.64$, $SD = 121.60$). In all of these samples, participants indicated their level of agreement with the HWK items using a 5-point Likert scale. The studies included in the meta-analyses are indicated by asterisks in the reference list. Fourteen of these 17 independent samples have been published in 12 separate articles (1 article contains 3 independent samples).

Thirteen of the studies were experimental in nature and 4 were correlational. The experimental studies included manipulations of goal difficulty, goal specificity, goal publicness, incentives, and framing effects. Eleven of the studies used students in laboratory settings, 5 employed students in natural settings (i.e., performing academic tasks), and 1 study was based on employees in a field setting. It should be noted that included in these 17 samples are the 3 studies reported in DeShon and Landis (1997) and 1 of the 3 studies used in the Hollenbeck, Klein et al. (1989a) validation study. The Hollenbeck, Williams et al. (1989b) sample was excluded from the sample of studies so that the data used in the current analyses would be independent of those used in the original development of the scale.

Coding of Studies and Constructs

We did not obtain the full individual-level data sets. Rather we acquired the relevant item-level data for each of the 17 samples. Each study was coded for sample size, effects among the HWK scale items, and the item means and standard deviations. In addition, information was collected on each study regarding the three potential moderators mentioned earlier (task complexity, measurement timing, and goal origin). These variables were identified as potential moderators based on previous goal setting meta-analyses (Klein et al., 1999; Mento, Steel, & Karren, 1987; Wofford, Goodwin, & Premack, 1992; Wood, Mento, & Locke, 1987) and critiques of the HWK scale (DeShon & Landis 1997; Tubbs, 1993).

For task complexity, the studies were classified as employing low (e.g., perceptual speed, anagrams), moderate (e.g., scholastic course work), or high (e.g., multiple cue probability learning) complexity tasks based on the categories provided by Wood et al. (1987). With regard to measurement timing, studies were coded as to when goal commitment was assessed: prior to, during, or after task performance. For purposes of conducting the equivalence comparisons, studies assessing commitment after the task were combined with those assessing commitment during the task. This was done for two reasons. First, there were relatively few studies in these two groups. Second, posthoc reports of how committed one was to a goal are more likely to reflect commitment felt

toward the end of performing the task than the level of commitment prior to the task. Finally, for the examination of goal origin as a moderator, studies were coded as to whether the goals were self-set by or assigned to participants.

Meta-Analytic Procedures

Meta-analyses were conducted using the procedures outlined by Hunter and Schmidt (1990). Their procedures examine the degree to which correlation differences across studies can be accounted for by sampling error and other statistical artifacts. These procedures normally include corrections to account for unreliability in measurement to better estimate the true population correlation. However, since the unreliability for interitem correlations cannot be established, corrections for unreliability were not relevant for this study. Population correlations among each of the items in the HWK scale were estimated by using the observed correlations weighted for sample size and the calculated variance of correlations across samples. For the overall analysis, we performed an independent samples meta-analysis so that an individual's responses were not counted more than once. If participants completed the HWK measure more than once within the same study (e.g., for multiple trials or goals), interitem correlations were averaged within that study prior to calculating the weighted average across studies.

In addition to examining the equivalence of the scale across the potential moderators through the use of multisample confirmatory factor analysis presented below, the potential moderators were also analyzed using the meta-analysis moderation guidelines set forth by Hunter and Schmidt (1990) and Whitener (1990). In doing so, separate interitem correlation matrices were meta-analyzed for different subgroups (e.g., assigned versus self-set goals). Those metaanalyzed correlations were then examined in terms of the percentage of variance explained by artifacts and homogeneity significance tests. Credibility intervals were also examined for degree of spread and the inclusion of zero. Because the meta-analyses were conducted at the item level, the overall independent samples analysis and the moderator analyses, taken together, resulted in the calculation of over 300 separate meta-analyses.

Meta-Analysis and Confirmatory Factor Analysis

Most articles that combine meta-analysis and covariance structure models focus on testing a path model because that is the only analysis available given the results from a typical meta-analysis. Therefore, most of the existing written guidelines apply to those procedures (readers are referred to Viswesvaran and Ones (1995) and Shadish (1996) for detailed discussions). Many of these same issues apply when using confirmatory factor analyses to test a measurement model and we have attempted to use the most widely accepted guidelines when appropriate. However, because this is the first measurement model meta-analysis that, to our knowledge, has been conducted, two procedures used in this study are worth noting. First, all of the confirmatory factor analyses

reported in this article were performed using covariance matrices and not a correlation matrix, which is typically what is obtained through the use of meta-analysis. Covariance matrices were created through the combination of both the observed correlations weighted for sample size and the sample size weighted item standard deviations. Using correlation matrices for covariance structure models is a debated issue and the problems with using correlations have been documented elsewhere (Bollen, 1989; Cudeck, 1989; Shadish, 1996).

Second, although most of the analyses performed have equal sample sizes for each estimated covariance, there were some analyses in which the meta-analyzed correlations had samples of different sizes. In those cases (analyses performed using items 4, 8, and 9), confirmatory models were given a sample size according to the harmonic mean (n_h) of the covariance matrix being analyzed, which is a rational approach discussed by Viswesvaran and Ones (1995). In no case did the use of the harmonic mean affect the significance of the parameter estimates, factor structure of the model, or model fit indices. In fact, in every circumstance when n_h was used, the same general results were found using the smallest cell size in the matrix.

Multisample Confirmatory Factor Analysis

To show that the proposed measurement items have identical relations across multiple situations, we sought to establish *measurement equivalence* across situational groups (e.g., Drasgow, 1984). It is important to note that to establish measurement equivalence there is no requirement of equal means or variances for either the individual items or the latent construct “goal commitment” across subgroups. We would fully expect that variance in the goal commitment factor and the item means would change across the different moderator subgroups proposed in this study (e.g., higher means and less variance in goal commitment for self-set goals as opposed to assigned goals).

In order to test for the presence of measurement equivalence across subgroups, we analyzed a series of two nested measurement models using a multisample analysis procedure in confirmatory factor analysis. In terms of the present study, multisample analysis analyzes data from each of our moderator groups simultaneously and ought to verify that a model, identical across the moderating factors, reproduces the sample data of each group to within sampling accuracy (Bentler, 1995). This multisample procedure allowed us to assess the fit of the proposed model by placing equality constraints on various parameter estimates. This procedure involved the use of two separate and distinct steps. The first step (or group of constraints) which we label *partial measurement equivalence* is a test of whether the *form* of the model is equivalent across groups. Model form equivalence is the simplest and least restrictive of all equivalence models performed in multisample analysis (Bollen, 1989). If responses to the proposed goal commitment construct do not hold to the same conceptual framework (i.e., unidimensionality), it makes little sense to examine for similar calibration among the measures (Vandenberg & Self, 1993). In the case of our model, acceptable fit statistics for a model of partial measurement

equivalence will help to show unidimensionality of the goal commitment measure across levels of goal origin, measurement timing, and task complexity. The second step, which we label *full measurement equivalence,* involves placing equality constraints on the item factor loadings across subgroups. By specifying invariant factor loadings across samples we are testing whether the subgroups calibrate the items in a similar fashion across situations (Riordan & Vandenberg, 1994). In other words, acceptable model fit statistics at this step will help to show scaling unit equality (Schaubroeck & Green, 1989; Vandenberg & Self, 1993) across levels of goal origin, measurement timing, and task complexity.

For both our multisample analyses and initial scale refinement analyses, we used various statistical indices to help determine model fit. One of the most popular and often used statistics to assess model fit is the χ^2 statistic. The fact that χ^2 is largely affected by sample size (such that large samples lead toward rejection of a model even if the model has reasonable levels of fit) has been well documented (Bentler & Bonett, 1980; Hu & Bentler, 1995). Obviously, given the large aggregate sample used here, the χ^2 goodness of fit test was highly significant throughout this study and is provided mainly for completeness. However, many other practical fit indices have been developed that help to overcome the limitations of the χ^2 statistic and the use of multiple indices of fit, each with different assumptions, to evaluate model fit has been recommended (Bollen, 1990). We chose to use a number of popular indices, including the goodness of fit index (GFI; Jöreskog & Sörbom, 1984), the non-normed fit index (NNFI; Bentler & Bonett, 1980), the comparative fit index (CFI; Bentler, 1990) and the root mean squared error of approximation (RMSEA; Steiger, 1990). Conventionally, the GFI, NNFI, and CFI show acceptable levels of fit when their values exceed .90. The RMSEA shows acceptable levels of fit when its values are below .08 (Brown & Cudeck, 1993).

RESULTS

Results of the independent samples meta-analysis are presented in Table 2. This table presents the weighted sample means and standard deviations for each item as well as the meta-analyzed interitem correlations.

Initial Analyses

As an initial analysis, all nine HWK items were regressed on a single latent variable using confirmatory factor analysis. The results of this analysis are presented in Table 3. The confirmatory factor analysis indicated that the unidimensional model using all nine items marginally represented the observed covariances between the HWK scale items. An exploratory factor analysis (principal component with varimax rotation) was also conducted on the nine-item correlation matrix. Similar to some previous findings, the results revealed a two-factor solution based on the scree plot and Kaiser criterion. Also consistent with previous findings, the second factor accounted for a relatively small

TABLE 2

Weighted Sample Results for the Nine-Item HWK Measure across Independent Samples

					Interitem correlations (Item No.)								
Item	N^a	K	Mean	SD	1	2	3	4	5	6	7	8	9
1	2918	17	3.58	0.96	.54[b]								
2	2918	17	3.70	0.91	.40 (.12)[c]	.42							
3	2918	17	2.97	0.97	.31 (.13)	.32 (.15)	.30						
4	2748	17	3.69	0.91	.39 (.11)	.27 (.08)	.18 (.11)	.56					
5	2918	17	3.61	0.91	.30 (.12)	.22 (.08)	.13 (.06)	.42 (.15)	.57				
6	2918	17	3.50	0.94	.36 (.06)	.30 (.09)	.23 (.08)	.44 (.13)	.41 (.10)	.56			
7	2918	17	3.67	0.86	.35 (.10)	.25 (.10)	.19 (.11)	.30 (.11)	.37 (.12)	.32 (.09)	.47		
8	1296	8	3.60	0.92	.23 (.10)	.17 (.08)	.06 (.00)	.38 (.00)	.58 (.05)	.33 (.06)	.27 (.08)	.46	
9	952	4	4.14	0.76	.31 (.00)	.19 (.00)	.12 (.06)	.40 (.00)	.35 (.00)	.35 (.02)	.27 (.00)	.32 (.03)	.46

Note. All meta-analyzed correlations are significant at the $p < .05$ level. Meta-analyzed correlations greater than 0.16 are significant at the $p < .01$ level.

[a] N is the aggregate sample size for each item; K is the number of studies in which the item was used. The actual N and K used to calculate each mean correlation is the lower of the two item N's and K's, except for the mean correlation between items 4 and 8, for which $N = 1125$ and $K = 6$.

[b] Item-total correlations based on the meta-analyzed matrix are on the principal diagonal.

[c] Standard deviation of the population correlations are in parentheses. Additional meta-analytic statistics are available upon request from the first author.

TABLE 3

Metaconfirmatory Factor Analysis Results of the HWK Scale

Item	Full nine-item model[a] Factor loading[b]	Full nine-item model[a] R^2	Proposed five-item model[a] Factor loading[b]	Proposed five-item model[a] R^2
1	.569	.324	.563	.317
2	.436	.190		
3	.312	.098		
4	.654	.428	.660	.435
5	.673	.452	.626	.392
6	.633	.400	.649	.421
7	.522	.272	.531	.282
8	.579	.335		
9	.538	.289		
$\chi^2(df)$	533.31 (27)		68.52 (5)	
NNFI	0.807		0.954	
RMSEA	0.105		0.067	
GFI	0.925		0.991	
CFI	0.855		0.977	
α	0.792		0.743	

Note. CFI, comparative fit index; GFI, goodness of fit index; RMSEA, root mean square error of approximation; NNFI, non-normed fit index.

[a] n_h (9-item) = 1695; n_h = (5-item) = 2868.

[b] All factor loadings are significant at the $p < .01$ level.

amount of the variance explained (14%). The observed reliability (α) for the nine-item scale was 0.79.

Scale Refinement

The primary purpose of this study concerned the identification of a robust, unidimensional measure of goal commitment from the nine HWK scale items. Since the full nine items yielded less than desirable fit statistics for a unidimensional model, efforts were made to discern a subset of items that would produce acceptable fit statistics for a unidimensional model and have other desirable psychometric properties. The process used in refining the scale revolved around three key considerations or decision points. While our approach to refining the scale was empirical, our first consideration was conceptual. In examining the various scales that could be created from the nine HWK scale items, the item "I am strongly committed to pursuing this goal" was used as the theoretical core item. As such, it was felt that this item should be retained in the resulting scale and that other items should be evaluated, in part, relative to this item. Second, the statistics obtained from confirmatory factor analyses (including factor loadings, standardized residuals, and model fit statistics) were considered. These statistics were used to determine the relative appropriateness, from an empirical perspective, of various possible scales. Finally, we took into

consideration the reliabilities of the resulting scales. The application of these decision points in refining the scale is described below.

Given that the nine-item model yielded less than desirable levels of fit, we examined the correlation matrices, factor loadings, and residuals from that full model. Those statistics indicated that items 2 and 3 were the poorest items in that they produced the lowest factor loadings, highest residuals, and reduced overall model fit the most compared to the other seven items. The fit of the resulting seven-item model was then examined and was also judged to be less then desirable. The same process was again used to identify additional items for possible elimination. Of the remaining items, items 8 and 9 were the most inconsistent, lowering the overall fit when they were included in various models. The resulting five-item scale, described in more detail below, produced highly acceptable levels of fit. While even higher levels of overall model fit statistics could be achieved by dropping additional items, further analyses indicated that such gains in model fit come at the expense of scale reliability. We thus concluded that the five-item measure, consisting of items 1, 4, 5, 6, and 7, is the most appropriate scale for future researchers to use when measuring goal commitment. These items are italicized in Table 1.

The results of a confirmatory factor analysis of these five items are presented in Table 3. The GFI, NNFI, CFI, and RMSEA fit indices all reached highly acceptable levels of fit and better levels of fit than the full nine-item model. The reduced scale produced an α reliability of 0.74, only slightly lower than that observed with the full nine items. An exploratory factor analysis was also performed and those results also supported a one-factor solution. Examination of the scree plot and the Kaiser criterion both denoted a one-factor solution with that factor accounting for 49% of the variance. Factor loadings for the five-item measure ranged from .65 to .74 and there were no standardized residuals above the .05 level. The primary purpose of this study was thus achieved as a clearly unidimensional measure of goal commitment was derived from the nine HWK scale items.

It should be noted that the items removed through this process are largely consistent with previous suggestions. Hollenbeck, Klein et al. (1989a) suggested the elimination of items 8 and 9, a recommendation supported by the current analyses. Items 2 and 3, the other items eliminated through the current analyses, are two of the three items that DeShon and Landis (1997) suggested removing. Therefore, the empirical approach taken here, which concludes that a five-item scale best represents the underlying construct of goal commitment, is largely consistent with and addresses the dimensionality concerns based on previous subjective categorizations of the items. This study does differ from the conclusions of DeShon and Landis (1997) in that the first item, “It’s hard to take this goal seriously,” is retained as part of the commitment scale. They interpreted this item as assessing performance expectancies. While expectancies might influence how seriously one takes a goal, any number of other suggested antecedents of goal commitment (e.g., incentives, task experience, social influence) could also influence how one would respond to this item regardless of expectancies.

Stability of Scale Structure

Three research questions were examined concerning the equivalence of the derived five-item scale across a variety of settings and contexts (task complexity, measurement timing, and goal origin). To examine the effects of these potential moderator variables on the dimensionality of the five-item measure, we performed multisample confirmatory factor analysis as well as typical meta-analysis moderator analyses to compare results across subgroups of studies. Multisample confirmatory factor analyses progressed by testing two separate models for each moderator subgroup, a model specifying *partial measurement equivalence* (model form constrained to be equal across samples) and a model specifying *full measurement equivalence* (model form and item factor loadings constrained to be equal across samples).

Task complexity. Multisample confirmatory factor analyses results comparing groups of low ($n = 1195$), moderate ($n = 1396$), and high ($n = 273$) complexity samples and their responses to the recommended five-item goal commitment scale are provided in Table 4. The partial measurement equivalence model tested the equality of factor form across samples. The χ^2 statistic for the model was significant, which was expected given the large sample sizes involved. More practical fit indices showed high levels of model fit across all three complexity levels. Thus we find support for partial measurement equivalence, indicating unidimensionality of the goal commitment scale across levels of task complexity

The results when examining the more restrictive model of full measurement equivalence, where factor loadings were constrained to be equal, were also indicative of good model fit. Although the χ^2 statistic was significant, all other practical fit indices showed acceptable levels of model fit. In addition, these fit indices did not drop substantially between the testing of partial and full measurement equivalence. In addition to examining overall model fit statistics, it is possible to examine an indicator of how significant each constraint was in testing model fit through the use of modification indices. Modification indices allow researchers to examine how model fit statistics would change given the release of any given constraint imposed on the model.

Using the modification indices for releasing constraints in this multisample analysis showed that some of the paths constrained to be equal were significant from a purely statistical standpoint (i.e., χ^2). However, when dealing with large sample sizes, both the size of the statistic and the actual parameter differences should be examined (e.g., Kaplan, 1990, 1995). If the χ^2 statistic is high but parameter differences are low, the result of the statistical test can be seen as being mainly due to sample size rather than some fundamental internal specification error. In this case, the χ^2 values that were significant were still fairly small given the sample size and the parameter differences between groups were small and showed no discernable pattern. Therefore we conclude that the proposed goal commitment measure is equivalent across levels of low, moderate, and high task complexity.

TABLE 4

Tests Assessing the Measurement Equivalence of the Hypothesized Model for Goal Commitment across Various Subgroups

Model Specifications across subgroups	*df*	χ^2	Model comparison	Δ *df*	$\Delta \chi^2$	GFI	CFI	NNFI	RMSEA
Survey timing—Prior (n = 2716; K = 13) vs. during/after (n = 1200; K = 8)									
Partial measurement equivalence	10	82.72				.99	.98	.96	.04
Full measurement equivalence	15	100.72	Partial vs. full	5	18.00	.99	.98	.96	.04
Goal origin—Assigned (n = 1782; K = 11) vs. self-set (n = 1308; K = 7)									
Partial measurement equivalence	10	141.18				.98	.96	.92	.06
Full measurement equivalence	15	171.73	Partial vs. full	5	30.55	.98	.95	.92	.06
Task complexity—Low (n = 1195; K = 7) vs. moderate (n = 1396; K = 7) vs. high (n = 273; K = 3)									
Partial measurement equivalence	15	97.17				.99	.97	.94	.04
Full measurement equivalence	25	166.66	Partial vs. full	10	69.49	.98	.95	.94	.05

Note. CFI, comparative fit index; GFI, goodness of fit index; RMSEA, root mean square error of approximation; NNFI, non-normed fit index.

[a] *Partial measurement equivalence* is a model in which no restrictions other than the form of the model are imposed across subgroups.

[b] *Full measurement equivalence* is the same model as partial measurement equivalence with the addition of equality constraints on the factor loadings for each item.

In addition to the multisample confirmatory factor analysis we also performed more traditional meta-analytical moderation tests. Differences between the meta-analyzed correlations for the low, moderate, and high complexity samples were found to be minimal based on both the degree of overlap of the confidence intervals for each correlation and the percentages of variance explained in correlation distributions leading us to the same conclusion of measurement equivalence across levels of task complexity.

Measurement timing. Multisample confirmatory factor analysis results for the prior (n = 2716) and during/after (n = 1200) subsamples are also found in Table 4. The results for both full and partial measurement equivalence were similar in nature to the results for task complexity. Results for the model testing partial measurement equivalence suggests that dimensionality is not affected by the timing of the measurement given the high levels of model fit indicated by the practical fit indices. Constraining the factor loadings to be equal across samples in order to show full measurement equivalence also produced high levels of fit statistics and there was no discernable drop in these indices from the previous model. Modification indices were similar in nature to those for task complexity and gave no indication that releasing the factor loading constraints would significantly improve model fit. From a meta-analytical moderation perspective, minimal differences were observed between the two subsample correlation matrices. We therefore conclude that the equivalence of the proposed five-item measure of goal commitment is not affected by measurement timing.

Goal origin. Results from the multisample confirmatory factor analysis testing differences between the self-set goal sample (n = 1308) and the assigned goal sample (n = 1782) are also provided in Table 4. Model fit results for the partial measurement equivalence model suggest that dimensionality is not affected by goal origin. Full measurement equivalence is also supported both by high levels of model fit and no worsening of model fit when factor loading constraints are imposed. Results similar to the preceding meta-analytical moderation analyses for task complexity were found when performed for goal origin. All the measures of moderation for the interitem correlations of the proposed five-item scale provided no pattern and no hints that the correlations would be any different than that which would be expected by chance. We therefore conclude that goal origin does not significantly affect measurement equivalence across the two samples.

Scale Level and Variation

As noted earlier, while the structure of the scale should be stable across settings and contexts, differences in scale level or variation can be expected and do not invalidate the scale. Scale descriptive statistics as well as the α reliabilities for each of the subgroups from each of the moderator analyses were examined to identify the impact different settings and contexts have on the scale and to provide normative information for future researchers. Given

the sample sizes involved, differences between virtually every pair of means were statistically significant. In order to evaluate the practical significance of these differences, effect sizes between means were calculated and categorized as weak, moderate, or strong based on Cohen and Cohen's (1983) standards.

Across the task complexity subgroups, there appears to be a curvilinear relationship with respect to the mean level of commitment reported. Commitment was highest for moderately complex tasks ($M = 3.92$; $SD = 0.58$) and lowest for high complexity tasks ($M = 2.62$; $SD = 0.69$). The differences in mean commitment level between the low ($M = 3.47$; $SD = 0.71$) and high task complexity studies and those between the moderate and high task complexity studies are both strong effects. The difference in mean commitment between the low and moderate task complexity studies is a moderate effect. It should be noted that potential differences in other important variables (e.g., ability, goal difficulty) are not accounted for in these comparisons. The scale descriptive statistics were substantially similar for the two measurement timing subgroups. The difference in mean commitment between studies assessing commitment prior to the task ($M = 3.67$; $SD = 0.62$) versus those assessing commitment during and after the task ($M = 3.54$; $SD = 0.63$) constituted a weak effect. There were notable differences in the scale descriptive statistics across the goal origin subgroups. As could be expected, commitment to self-set goals was substantially higher ($M = 3.99$; $SD = 0.50$) than commitment to assigned goals ($M = 3.25$; $SD = 0.76$) across the included studies. That difference is a strong effect, although it should again be noted that potential influences of other variables have not been fully controlled. As also noted above, these scale level differences do not invalidate the scale as mean differences should be expected across theoretically relevant contexts or settings.

In terms of variability, the differences between subgroups, both for task complexity and goal origin, appear to be largely a function of ceiling effects. As commitment levels become higher, variability is reduced due to the upper limit of the response scale. There was not much difference in the observed alpha reliability estimates between the task complexity subgroups (.73 for moderate complexity, .76 for low complexity, and .77 for high complexity) compared to the overall sample (.74) and the reliabilities for the two measurement timing subgroups were identical (.73 for each). The largest difference in reliability was evident for the goal origin subgroups where the α reliability was lower for self-set (.70) goals than for assigned (.78) goals. This reduced reliability is most likely due to the restricted variance observed in the scale for self-set goals. However, the structure of the scale remained unchanged across goal origin as indicated by the multisample confirmatory factor analyses and the meta-analytical moderation analyses. Therefore, the five-item scale can appropriately be used across a variety of conditions and settings even though lower reliabilities may be obtained for self-set goals.

Follow-Up Assessment of Construct Relevance

To complement the current empirical refinement the HWK scale and to confirm some of the subjective judgments made regarding the scale items, a

follow-up examination of the construct relevance of the HWK items was conducted. All of the Management and Industrial/Organizational Psychology doctoral students at two institutions were given the conceptual definition of the goal commitment construct and 18 randomly arranged items. These items included all 9 HWK items along with 9 items reflecting other related constructs (subjective goal difficulty, goal specificity, and performance expectancies). These expert judges were asked to rate the extent to which each item reflected the definition of goal commitment using a 5-point Likert scale. Twenty-nine students responded, constituting a 63% response rate.

The average rating given to the 9 HWK items ($M = 3.84$) was significantly higher then the average rating given to the 9 noncommitment items ($M = 1.84$; $t = 20.98$, $p < .01$). Furthermore, the items that make up the proposed 5-item scale ($M = 4.08$) were rated significantly higher than the 4 items we are suggesting be eliminated from the scale ($M = 3.53$; $t = 5.04$, $p < .01$). The opinions of the expert judges thus support the conclusion that the proposed 5-item scale is construct relevant and more cleanly reflects the construct then the full 9-item scale. The experts also confirmed our assumption that the item "I am strongly committed to pursuing this goal" is the theoretical core item, as this item received an unanimous rating of "5."

It should be noted that one of the items excluded from the scale, "I am willing to put forth a great deal of effort beyond what I'd normally do to achieve this goal," was also rated as highly construct relevant by the expert judges. This is a long, complex item that is basically a difference score with an unmeasured referent (i.e., it taps the difference between how hard the person is willing to work on a goal compared to what he or she would normally do). The previously reported empirical analyses did not support the inclusion of this item. While the expert judges were able to work through the length and complexity of the item to see it as construct relevant, the item appears to cause some confusion for the typical respondent (naive subjects). While the inclusion of this item in the scale would boost the coefficient α slightly (.03), doing so substantially reduces overall model fit. Construct relevance is a necessary but not sufficient reason for item inclusion. Given the additional concerns of parsimony and item clarity, we do not feel that this item should be included as part of the optimal scale.

DISCUSSION

The purpose of this study was to address the conflicting recommendations in the literature regarding what version of the HWK scale, if any, should be used to assess goal commitment. The current study used item-level meta-analyses to combine data from multiple samples, creating a covariance matrix based on 2918 subjects in order to minimize the potential bias of sampling error. That covariance matrix was used for both confirmatory and exploratory factor analyses in order to examine the dimensionality of the HWK scale. The unidimensional factor structure of the nine-item scale was not fully supported so steps were undertaken to discern a subset of items that would produce

acceptable fit statistics for a unidimensional model and have other desirable psychometric properties.

The results indicate that a five-item scale (items 1, 4, 5, 6, and 7) best represents the unidimensional construct of goal commitment. Using these five items provides a high degree of fit as a unidimensional construct and promises the highest level of practical significance. Multiple sample confirmatory factor analyses then examined the equivalence of that derived scale across studies with respect to measurement timing, goal origin, and task complexity. For all three of these potential moderators, the proposed five-item scale was equivalent and unidimensional across subgroups of studies. The proposed five-item scale thus appears to be highly stable and robust across a variety of settings and contexts. The proposed five-item scale was also judged to be construct relevant and more strongly so than the excluded items.

This effort reflects the first combined use of meta-analytic and confirmatory factor analytic techniques to test a measurement model. This allowed for the creation of a database substantially larger than previously factor-analyzed samples containing these scale items and the reduction of sampling error as an alternative explanation for discrepant findings in those samples. The results clarified a number of debated issues that had arisen out of previous small sample factor analyses. The main deterrents to using this approach are the large number of separate meta-analyses which need to be performed and the fact that very few studies report interitem correlations limiting the availability of the information needed for such an analysis.

This study is fairly unique in another respect. While not as involved as the collaborative effort of Latham, Erez, and Locke (1988), this study does represent the joint resolution of a scientific dispute. This study is the result of authors previously supporting the unidimensionality of the HWK scale joining with an author who had argued that the scale was multidimensional to share data and ideas in order to resolve the disagreement. Finally, the follow-up assessment of item relevance revealed an interesting difference in the way naive subjects and expert judges responded to one of the excluded items. This raises some interesting questions about the appropriate use of expert judges in the validation process which are deserving of additional research.

Limitations of the Study

Problems with statistical power and meta-analysis have been discussed elsewhere (Hunter & Schmidt, 1990; Sackett, Harris, & Orr, 1986) and one limitation of this study was our inability to conduct a hierarchical breakdown of moderators because of the limited number of independent samples in each subset. Both meta-analysis and the use of SEM call for many subjective decisions (e.g., Wanous, Sullivan, & Malinak, 1989; Viswesvaran & Ones, 1995) and although many of the usual subjective decisions were not applicable due to our focused investigation some subjective decisions were made regarding moderator classifications and the significance of model fit. Although the potential for bias needs to be recognized, we attempted to make decisions as objectively as possible.

One such subjective decision was the procedure used to identify the studies that were included in this measurement model meta-analysis. As described earlier, accepted procedures for systematically identifying studies meeting the inclusion criteria could not be used. Instead, three different researchers who had published studies using the HWK measure were asked to provide the relevant statistics from all available data sets meeting the inclusion criteria. This led to a sample of studies that were both published and unpublished and represented a variety of samples, settings, and contexts. The studies also came from authors who had previously supported as well as argued against the unidimensionality of the HWK scale. It could be asserted, however, that the choice of authors or the failure to include studies conducted by a broader set of authors might have biased the results. It should be noted, however, that given the aggregate sample size used in this study, a considerable number of additional respondents would be needed to have any impact on the reported results. Furthermore, those additional respondents would have to react to these same scale items in a substantially different manner in order to alter the findings reported here. Therefore, while the selection of studies for inclusion may have been unusual for a meta-analysis, one can have a high degree of confidence in the stability and generalizability of the results.

Conclusions

Goal commitment is an essential moderator of the linkage between goals and behavior. Given the theoretical centrality of goal commitment, it is critical to have a valid measure of this construct. The full nine-item HWK scale was valuable in pulling together previously used single item measures of commitment and allowing the accumulation of research on goal commitment based on a common measure. However, the five-item measure presented here is purer and more efficient than the full nine-item HWK scale. The same cannot be said for the four-item scale Hollenbeck, Klein, et al. (1989a) suggested be used when only a manipulation check is necessary (items 1–4). The use of that four-item scale can no longer be recommended given that (a) the four-item scale shares only two items with the proposed five-item scale and (b) the limited gain in efficiency from using a four- versus a five-item scale. The evidence presented in this study suggests that the proposed five-item scale is a psychometrically sound, construct relevant, robust, and widely generalizable measure of one's determination to reach a goal. It is therefore recommended that this five-item scale be used in future research requiring a self-report measure of goal commitment.

REFERENCES

Austin, J. T., & Vancouver, J. B. (1996). Goal constructs in psychology: Process and content. *Psychological Bulletin,* 120, 338–375.

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin,* 107, 238–246.

Bentler, P. M. (1995). *EQS structural equations program manual.* Encino, CA: Multivariate Software, Inc.

Bentler, P. M., & Bonnet, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin,* 88, 588–606.

Bollen, P. M. (1989). *Structural equations with latent variables.* New York: Wiley.

Bollen, P. M. (1990). Overall fit in covariance structure models: Two types of sample size effects. *Psychological Bulletin,* 107, 256–259.

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

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin,* 105, 317–327.

*DeShon, R., & Landis, R. S. (1997). The dimensionality of the Hollenbeck, Williams, and Klein (1989) measure of goal commitment on complex tasks. *Organizational Behavior and Human Decision Processes,* 70, 105–116.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent reactions with external variables are central issues. *Psychological Bulletin,* 95, 134–135.

Earley, P. C., & Kanfer, R. (1985). The influence of component participation and role models on goal acceptance, goal satisfaction, and performance. *Organizational Behavior and Human Decision Processes,* 36, 378–390.

Hollenbeck, J. R., & Klein, H. J. (1987). Goal commitment and the goal-setting process: Problems, prospects, and proposals for future research. *Journal of Applied Psychology,* 72, 212–220.

Hollenbeck, J. R., Klein, H. J., O'Leary, A. M., & Wright, P.M. (1989a). Investigation of the construct validity of a self-report measure of goal commitment. *Journal of Applied Psychology,* 74, 951–956.

Hollenbeck, J. R., Williams, C. L., & Klein, H. J. (1989b). An empirical examination of the antecedents of commitment to difficult goals. *Journal of Applied Psychology,* 74, 18–23.

Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In Hoyle, R.H. (Ed.), *Structural equations modeling: concepts, issues and applications.* Thousand Oaks, CA: Sage.

Hunter, J. E., & Schmidt, G. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Jöreskog, K., & Sörbom, D. (1986). *LISREL 6: Analysis of linear structural relationships by maximum likelihood and least squares methods.* Mooresville, IN: Scientific Software.

Kaplan, D. (1990). Evaluating and modifying covariance structure models: A review and recommendation. *Multivariate Behavioral Research,* 28, 483–498.

Kaplan, D. (1995). Statistical power in structural equation modeling. In Hoyle, R.H. (Ed.) *Structural equations modeling: Concepts, issues and applications.* Thousand Oaks, CA: Sage.

*Klein, H. J. (1991). Further evidence on the relationship between goal setting and expectancy theories. *Organizational Behavior and Human Decision Processes,* 49, 230–257.

*Klein, H. J. (1994). *Goal specificity and performance: Is it level or variation?* Unpublished manuscript, Ohio State University.

*Klein, H. J. (1996) *The meaning and predictive efficacy of personal goals.* Unpublished manuscript, Ohio State University.

*Klein, H. J., & Kim, J. S. (1998). A field study of the influence of situational constraints, leader-member exchange and goal commitment on performance. *Academy of Management Journal,* 41, 88–95.

Klein, H. J., Wesson, M. J., Hollenbeck, J. R., & Alge, B.J. (1999). Goal commitment and the goal setting process: Conceptual clarification and empirical synthesis. *Journal of Applied Psychology,* 84, 885–896.

*Klein, H. J., Whitener, E. M., & Ilgen, D. R. (1990). The role of goal specificity in the goal setting process. *Motivation and Emotion,* 14, 179–193.

*Klein, H. J., & Wright, P. M. (1994). Antecedents of goal commitment: An empirical examination of personal and situational determinants. *Journal of Applied Social Psychology,* 24, 95–114.

*Lane, M. C., Sterling, C. S., & Klein, H. J. (1992). *Motivation and commitment in the transfer of training process.* Unpublished manuscript, Ohio State University.

Latham, G. P., Erez, M., & Locke, E. A. (1988). Resolving scientific disputes by the joint design of crucial experiments by the antagonists: Application to the Erez-Latham dispute regarding participation in goal setting. *Journal of Applied Psychology Monograph,* 73, 753–772.

Locke, E. A., & Latham G. P. (1990). *A theory of goal setting and task performance.* Englewood Cliffs, NJ: Prentice Hall International..

Locke, E. A., Latham, G. P., & Erez, M. (1988). The determinants of goal acceptance and commitment. *Academy of Management Review,* 13, 23–39.

*Mento, A. J., Locke, E. A., & Klein, H. J. (1992). The relationship of goal level to valence and instrumentality. *Journal of Applied Psychology,* 77, 395–405.

Mento, A. J., Steel, R. P., & Karren, R. J. (1987). A meta-analytic study of the effects of goal setting on task performance: 1966–1984. *Organizational Behavior and Human Decision Processes,* 39, 52–83.

Riordan, C. M., & Vandenberg, R. J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? *Journal of Management,* 20, 643–671.

Sackett, P. R., Harris, M. M., & Orr, J. M. (1986). On seeking moderator variables in the meta-analysis of correlational data: A monte-carlo investigation of statistical power and resistance to Type-1 error. *Journal of Applied Psychology,* 71, 302–310.

Schaubroeck, J., & Green, S. G. (1989). Confirmatory factor analytic procedures for assessing change during organizational entry. *Journal of Applied Psychology,* 74, 892–900.

Shadish, W. R. (1996). Meta-analysis and the exploration of causal mediating processes: A primer of examples, methods, and issues. *Psychological Methods,* 1, 47–65.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research,* 25, 173–180.

Tubbs, M. E. (1993). Commitment as a moderator of the goal-performance relation: A case for clearer construct definition. *Journal of Applied Psychology,* 78, 86–97.

Tubbs, M. E., & Dahl, J. G. (1991). An empirical comparison of self-report and discrepancy measures of goal commitment. *Journal of Applied Psychology,* 76, 708–716.

Tubbs, M. E., & Ekeberg, S. E. (1991). The role of intentions in work motivation: Implications for goal-setting theory and research. *Academy of Management Review,* 16, 180–199.

Vandenberg, R. J., & Self, R. M. (1993). Assessing newcomers' changing commitments to the organization during the first 6 months of work. *Journal of Applied Psychology,* 78, 557–568.

Viswesvaran, C., & Ones, D. S. (1995). Theory testing: Combining psychometric meta-analysis and structural equations modeling. *Personnel Psychology,* 48, 865–885.

Wanous, J. P., Sullivan, S. E., & Malinak, J. (1989). The role of judgment calls in meta-analysis. *Journal of Applied Psychology,* 74, 259–264.

Whitener, E. M. (1990). Confusion of confidence intervals and credibility intervals in meta-analysis. *Journal of Applied Psychology,* 75, 315–321.

Wofford, J. C., Goodwin, V. L., & Premack, S. (1992). Meta-analysis of the antecedents of personal goal level and of the antecedents and consequences of goal commitment. *Journal of Management,* 18, 595–615.

Wood, R. E., Mento, A. J., & Locke, E. A. (1987). Task complexity as a moderator of goal effects: A meta-analysis. *Journal of Applied Psychology,* 72, 416–425.

*Wright, P. M. (1992). An examination of the relationships among monetary incentives, goal level, goal commitment, and performance. *Journal of Management,* 18, 677–693.

*Wright, P. M., George, J. M., Farnsworth, S. R., & McMahan, G. C. (1993). Productivity and extra-role behavior: The effects of goals and incentives on spontaneous helping. *Journal of Applied Psychology,* 78, 374–381.

*Wright, P. M., Hollenbeck, J. R., Wolf, S., & McMahan, G. C. (1995). The effects of varying goal difficulty operationalizations on goal setting outcomes and processes. *Organizational Behavior and Human Decision Processes,* 61, 28–43.

*Wright, P. M., & Kacmar, M. K. (1994). Goal specificity as a determinant of goal commitment and goal change. *Organizational Behavior and Human Decision Processes,* 59, 242–260.

*Wright, P. M., & Kacmar, K. M. (1995). Mediating roles of self-set goals, goal commitment, self-efficacy, and attractiveness in the incentive-performance relation. *Human Performance,* 8, 263–296.

*Wright, P. M., O'Leary-Kelly, A. M., Cortina, J. M., Klein, H. J., & Hollenbeck, J. R. (1994). On the meaning and measurement of goal commitment. *Journal of Applied Psychology,* 79, 795–803.

Received April 19, 1999; published online February 8, 2001