Stages of Psychometric Measure Development: The example of the Generalized Expertise Measure (GEM)

Marie-Line Germain
Barry University & City College

This paper chronicles the steps, methods, and presents hypothetical results of quantitative and qualitative studies being conducted to develop a Generalized Expertise Measure (GEM). Per Hinkin (1995), the stages of scale development are domain and item generation, content expert validation, and pilot test. Content/face validity and internal consistency of the scores of the GEM are discussed, as well as directions to ensure that the psychometric properties of the scale are theoretically and empirically sound.

Keywords: Measure, Research Methods, Expertise

Since its humble beginnings in the late 1950s, expertise has slowly permeated the human resource development literature. This seemingly obscure concept became an intriguing research topic as artificial intelligence and cognitive psychology further developed in the mid- to late sixties. Swanson & Holton (2001) define expertise as a "displayed behavior within a specialized domain and / or related domain in the form of consistently demonstrated actions of an individual that are both optimally efficient in their execution and effective in their results." (p. 241).

The absence of empirical evidence may well be the main reason of the gradual development of understanding of expertise in the last three decades (cf. Bédard & Chi, 1992). The meticulous standards of scientific methodologies invalidate the pervasive existence of anecdotal evidence in the popular press of expertise in organizational settings. The past 15 years, however, have seen an upsurge in the pace of expertise research, as indicated in the growing number of peer-reviewed publications in the area (Holton & Swanson, 2001). There are almost as many definitions of "experts" as there are researchers who study them (Hoffman, Shadbolt, Burton, and Klein, 1995). However, some of these conceptual research studies have identified various common themes or dimensions associated with expertise, namely knowledge, experience in the field, and problem-solving skills, between others.

In order to develop expertise into a more substantial construct and theory than what exists presently, the concept needs to be further empirically tested and validated. Rigorous qualitative and quantitative research studies on the constructs of expertise are particularly critical, and supposedly are the only logical step to draw the concepts into an intelligible whole. Otherwise, a sound understanding of expertise would remain virtually non-existent. Schwab (1980) argues that adequate measurement is required to make theoretical progress possible. The purpose of this study, therefore, is to show the developmental steps of a measurement scale of expertise. The main research question is whether it is possible to develop a scale that truly measure expertise.

Stages of Scale Development

The American Psychological Association (1985, as quoted in Hinkin, 1995) established that sound measures must demonstrate content-validity, criterion-related validity, construct validity, and internal consistency. These criteria determine the psychometric validation of behavioral measures. Having closely examined 277 scale development practices in 75 studies, Hinkin (1995) argued that measures generally lack content validity in the item development stage and do not have strong and clear linkages with their theoretical domains. The current study addresses these two concerns by building content validity into the measure through the processes of domain identification, item generation, and judgment-quantification or content expert validation (DeVellis, 1991). The following sections outline the steps of scale development undertaken in this study to date: (1) Domain identification and item generation, (2) Content expert validation, and (3) Pilot test. The methodologies used were sequentially elaborated.

Domain Identification and Item Generation

The generation of items is the most important element of establishing sound measures (Hinkin, 1995). In this stage, the primary concern of the scale developer will be content validity. It is often viewed as the minimum psychometric requirement for measurement adequacy and is the first step in construct validation of a new measure (Schriesheim, Powers, Scandura, Gardiner, & Lankau, 1993). Content validity must be built into the measure through the development of items. As suggested by Schriesheim et al. (1993), content adequacy will be assessed

immediately after items have been developed as this will provide the opportunity to refine and / or replace items before preparing and administering a questionnaire. An inductive approach will be used, also called "grouping" or "classification from below" (Hunt, 1991). In an inductive scale development approach, there is little theory involved at the outset as we try to identify constructs and generate a measure from individual responses.

To generate themes and obtain more substantive insights pertinent to expertise, a first panel (Panel 1) will be formed, composed of six individuals who are generally considered experts in their field. Two non-probability sampling techniques, purposive and snowball were utilized in the selection of interview participants to ensure that they were "appropriate" opinion leaders with well-developed views on the research topic (Minichiello, Aroni, Timewell, & Alexander, 1995). Given the generative purpose of the interview, the sample size does not have to be large since "the validity, meaningfulness, and insights generated from qualitative inquiry have more to do with the information-richness of the cases selected and the observational/analytical capabilities of the researcher than with sample size" (Patton, 2002, p.185). The six panel members will first be contacted via e-mail and invited to participate in this study. The following e-mail will be sent: "We are in the process of developing a psychometric instrument that requires your expertise. Would you be willing to meet for a 90-minute discussion with a panel of 5 other members?" The goal of this panel is to discuss and define expertise.

At the beginning of the meeting, the researcher will ask them the following question: "What do you think expertise is, and according to you, what are the components of expertise?" The function of the researcher is to only facilitate the discussion. Each panel member will be given a chalk and will be able to write keywords or sentences on a board, and all panel members' duty will have to add to this brainstorming session. A semi-structured interview is appropriate for the current study since the existing limited information on expertise only allows for the development of flexible interview guides, not rigidly structured interview schedules (Miller & Crabtree, 1999b, p. 19).

Based on Panel 1's responses, the researcher will then write the questions. Interview transcripts and responses will be classified into categories by content analysis based on keywords and themes – This should be done by a team, which will constitute Panel 2. Members will sort the questions into content areas of expertise. Working as a team, Panel 2, also composed of six members, will go through this task quicker. We might also obtain a better structure at the end. The six members of this panel do not have to be "scholars". Anyone who works in a setting who has to consult with an expert to complete their work is eligible to be a member. Actually, the panel needs to match the expected target population. Therefore, "experts" (individuals with higher degrees) should not be included. While facilitating the meeting, it is important that the researcher removes herself from any debate.

Content analyses of interview transcripts will result in a certain number of themes (it is not atypical to have more than 30) which are perceived to be associated with expertise. These themes are generated from questions revolving around the perceived meaning of expertise by the Panel. In subsequent content analyses, these themes can be cross-checked with the literature review but this is not necessary. As a result, the themes can be conceptually grouped into sub-dimensions and dimensions. Table 1 shows an example of dimensions, sub-dimensions, and attributes associated with expertise.

Table 1. *Dimensions, Sub-Dimensions, and Behavioral Attributes Associated with Expertise*

D*imensions & Sub-dimensions*	*Examples of Behavioral Attributes*
Dimension 1 • Sub-dimension 1 • Sub-dimension 2 • Sub-dimension 3	Here, Write an example of a behavioral attribute associated with this sub-dimension.
Dimension 2 • Sub-dimension 1 • Sub-dimension 2 • Sub-dimension 3	
Dimension 3 • Sub-dimension 1 • Sub-dimension 2 • Sub-dimension 3	

Those generated items are subjected to a sorting process using the themes/construct elements by a third Panel (Panel 3). Panel 3 can be the same as Panel 2 in composition (both in the individuals and in member count).

The hypothetical 81 items are then subjected to content expert validation, which is a method for ensuring the content validity of the measurement instrument (Grant & Davis, 1997). It is essentially a sorting process which is used in this study to identify and delete theoretically incoherent items, and, thus, ensuring that the items in a scale demonstrate content adequacy (Hinkin, 1995). The Standard for Educational and Psychological Testing (American Educational Research Association, 1985) prescribed three criteria for expert panel members involved in content review process, namely relevant training, experience, and qualifications.

Following the suggestions of Grant and Davis (1997), the content experts will be asked to address three elements in examining the expertise instrument: representativeness, comprehensiveness, and clarity. Representativeness in this study refers to the degree to which each item reflects and operationalizes its nominated domain. To facilitate this evaluation process, the items are already categorized under their nominated domains prior to the evaluation process, and the definition of each of the identified domains is provided. The content experts are then asked to indicate the extent to which they perceive each individual item to be representative of the domain with which it was associated, by circling the most appropriate number in the 4-point rating scale (1= not representative, 2=minimally representative, 3=moderately representative, and 4= strongly representative). This first element forms the quantitative part of the content validation process. Hence, Panel 3 members will look at the questions and identify whether the questions capture the constructs, the closeness of the items to the constructs (Hinkin & Schriesheim, 1989). This will serve as a pre-test, allowing deletion of items. This panel's duty is to sort the items telling us "how much" each item measures each category. Here, it is not so much about deletion than it is about "no load".

The second task is to evaluate the comprehensiveness of the entire instrument by identifying items which they perceive to be incongruent with its nominated domain and, subsequently, assigning them to an alternative domain with which the items are better matched. Finally, the members are asked to identify the clarity of items construction and wording to ensure that there were no ambiguous and poorly written items. The last two elements form the qualitative parts of the evaluation process. Interrater consistency will be reported. Lawshe's (1975) Content Validity Ratio (CVR) can be utilized to assess the content expert judgment. The CVR value ranges from -1.00 and +1.00, where a CVR of 0.00 means that 50% of the experts in the panel believe that a measurement item is "essential" and, therefore, content valid. Lawshe's (1975) specified a formula for determining a minimum CVR for different panel sizes. According to this formula, a minimum CVR value of 0.49 is required for fifteen panel members. Using this procedure, 55 items with CVR value higher than 0.49 are retained in the scale, as shown in Table 2.

Categories will be defined and items will be written. Anchors are developed based on Panel 1's responses. The anchors do not have to be a Likert scale (Likert, 1932). Each construct should have about 15 questions.

Guided by the interview data, items (more than 80 is not exceptional, as shown in Table 2, which serves as an example with 81 items) are subsequently derived deductively and inductively, consistent with the definitions of each of the identified domains. This sequence is commonly utilized by researchers for theory development and item construction (for example, MacKenzie, Podsakoff, & Fetter, 1991; Mayfield, Mayfield, & Kopf, 1995; Podsakoff, MacKenzie, Moorman, & Fetter, 1990).

Table 2. *Example of CVR Computation Results*

Minimum CVR Value	*Number of Items Retained*	*Cumulative Number of Items*
0.89	5	5
0.79	6	11
0.69	29	40
0.59	15	55
0.49	6	61
0.39	4	65
0.29	10	75
0.19	3	78
0.09	2	80
0-0.09	0	80
<0	1	81

Scale Development and Pilot Study

Step 1: Design of the exploratory study. A pilot study of the items of the measure will then be conducted. The primary purpose of the pilot study is to measure the extent to which the instrument is able to "provide data of

sufficient quality and quantity to satisfy the objectives of the research" (Hunt, Sparkman, & Wilcox, 1982, p. 270). The generated items will be administered to a sample. This sample will be targeted towards the general population, but it is preferable to have employed and professional individuals. Also, per Schwab (1980), the item-to-response ratio should be close to 1:10 for each set of scales to be factor analyzed. Recent research, however, has found that in most cases, a sample size of 150 observations should be sufficient to obtain an accurate solution in exploratory factor analysis as long as item intercorrelations are reasonably strong (Guadagnoli & Velicer, 1988; Hinkin, 1995). Ultimately, we want the items to condense. The respondents are asked to evaluate their direct leader or supervisor with regard to their expertise using a 5-point Likert scale (Likert, 1932) (1= strongly disagree to 5= strongly agree).

Step 2: Reliability assessment. Reporting of internal consistency reliability is a necessary part of the scale development process (Hinkin, 1995). Reliability is a necessary pre-condition for validity (Nunnally, 1978). To assess the reliability of an instrument based on internal consistency, the minimum level of Cronbach's coefficient alpha (Price & Mueller, 1986) is .70 for basic research measures, following Nunnally's (1978) suggestion.

All descriptive item levels data including standard deviations (SD), correlations (r), reliability (α) and the mean (M) will also be reported for the factors of expertise, as shown in Table 3. All the factors intercorrelations will be calculated and should be less than 1.00 to be conceptually distinct. As for the reliability, the minimum acceptable should be .70, per Nunnally (1978). Nunnally & Bernstein (1994) also indicate that lower bound reliability should be at the minimum value of .70.

To allow precision in evaluating the new measure, as suggested by Hinkin (1995), confirmatory factor analyses (CFA) will be conducted using LISREL8 (Joreskog & Sorbom, 1993). LISREL is used to assess the quality of the factor structure by statistically testing the significance of the overall model and of item loadings on factors. Confirmatory factor analysis is a data reduction technique that assesses the interrelationships among a set of variables in an effort to find a new set of variables, fewer in number than the original set of variables, that expresses what is common among the original variables. Confirmatory factor analysis, unlike exploratory factor analysis, provides a complete and unified system for testing a priori models (Dillon & Goldstein, 1984). For confirmatory factor analysis, a minimum sample size of 200 has been recommended (Hoelter, 1983). It is in this step where *things* can go wrong. Indeed, Exploratory Factor Analysis (EFA) may be required depending on the results of the CFA. Exploratory factor analysis are undertaken to examine the factor structure of the scale (Tabachnick & Fidell, 1989). Principal component analysis generates factors. If the CFA confirms, then we will need to draw another sample and confirm again. If the CFA does not confirm, then EFA is used and a second sample drawn for a new CFA after item analysis. The overall purpose of exploratory and confirmatory factor analysis is to ensure the stability of the factor structure (Hinkin, 1995). Therefore, item deletions and revisions/modifications to the measurement can be expected on the basis of these analyses. Finally, a scree test can be calculated, which would indicate whether factors should be retained. Also, an oblique (Promax) rotation can be calculated, which would indicate whether the items which make up each single factor are or are not conceptually distinct.

The objective of the previous stages in the scale development process was to create measures that demonstrate validity and reliability (Hinkin, 1995). Construct validation is now essential to ensure the quality of the new measure (Schmitt & Klimoski, 1991). According to Cronbach & Meehl (1955), the demonstration of construct validity of a measure is the ultimate objective of the scale development. Campbell (1976) asserts that due to potential difficulties caused by common method variance, it is inappropriate to use the same sample both for scale development and for assessing construct validity. Also, the use of an independent sample to provide an application of the GEM will enhance its generalizability (Stone, 1978).

Table 3. *Means, Standard Deviations, Correlations and Reliabilities of Expertise Factors*

Expertise Factors	*M*	*SD*	*r*	*α*	*1*	*2*	*3*	*4*
Factor 1	-	-	-	-				
Factor 2	-	-	-	-	-			
Factor 3	-	-	-	-	-	-		
Factor 4	-	-	-	-	-	-	-	
Factor 5	-	-	-	-	-	-	-	-

As reported above, the first critical steps in the process are as follows: Interview data suggest that there are X dimensions and X sub-dimensions pertinent to expertise. The construct definitions guide the generation of X items that are subjected to content expert validation and pilot-tested. Subsequently, factor analyses are conducted to examine the psychometric properties of the scale. Throughout these stages, items can be deleted and refined to improve the reliability and validity of the scores on the scale. Following the content expert validation, the X items in

the initial pool can be reduced to X items. Data from the pilot test indicate whether these items are content valid and internally reliable. The high level of internal reliabilities among the X factors might suggest that the items that make up the sub-scale are measuring highly similar underlying attributes. The findings from principal component analysis can be conclusive or inconclusive and consistent or inconsistent with the hypothesized theoretically-derived factors. The items that make up each single factor confirm or disconfirm the initial a priori conceptualization. Confirmatory factor analysis can be conducted to further assess the construct validity of the measure. Additional item deletions and refinements can be expected in the next stage of confirmatory factor analysis. There are no typical issues related to the development process of the scale. It is important, however, to ensure that all panel members carefully follow the instructions provided by the scale developer. It is also important to make sure that the items resulting from Panel 1 are clear and non ambiguous. In order to ensure clarity the researcher may have them read and defined by a few randomly selected individuals for validation. The main pitfall of such an endeavor is that the items may not come out as representative of the concept of expertise after statistical analysis.

Conclusion and Contributions to the Field of HRD

The intent of this paper is to identify steps in the development and validation of a measure, using an expertise scale as an example. Validation of the scale should be done through data collection. Ideally, data from various fields could help in determining if the measure is indeed generalizable. Developing a scale that could measure expertise across a variety of fields could be of great help to Human Resource Development professionals. Such a scale could identify individuals that may or may not possess expert-like skills. The GEM may therefore be a useful tool for selection and hiring procedures.

References

American Educational Research Association, A. P. A., & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington DC: American Psychological Association.

Autry, J. A. (2001). *The servant leader*. Roseville, CA: Prima.

Bédard, J. & Chi, M. T. H. (1992). Expertise. *Current Directions in Psychological Science*, *1*(4), 135-139.

DeVellis, R. F. (1991). *Scale development: Theory and applications*. Newbury Park, CA: Sage Publications.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, *103*, 265-275.

Guertin, A. A., Guertin, W. H., & Ware, W. B. (1981). Distortion as a function of the number of factors rotated under varying levels of common variance and error. *Education and Psychological Measurement*, *41*, 1-9.

Hakstian, A. R., Roger, W. T., & Cattell, R. B. (1982). The behavior of number-of-factors rules with simulated data. *Multivariate Behavioral Research*, *17*, 193-219.

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management*, *21*(5), 967-988.

Hoelter, J. W. (1983). The analysis of covariance structures: Goodness-of-fit indices. *Sociological Methods and Research*, *11*, 325-344.

Hoffman, R. R., Shadbolt, N., Burton, A. M., & Klein, G. A. (1995). Eliciting knowledge from experts: A methodological analysis. *Organizational Behavior and Human Decision Processes*, *62*, 129-158.

Hunt, S. D., Sparkman, R. D., & Wilcox, J. B. (1982). The pretest in survey research: Issues and preliminary findings. *Journal of Marketing Research*, *14*, 269-273.

Joreskog, K., & Sorbom, D. (1993). *LISREL 8: Structural equation modeling with the simplis command language*. Hillsdale, NJ: Lawrence Erlbaum Associates.

King, N. (1994). The qualitative research interview. In C. Cassell & G. Symon (Eds.), *Qualitative methods in organizational research: A practical guide*. Pp. 14-36. Thousand Oaks, CA: Sage Publications.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*, 563-575.

Likert, R. (1932). A Technique for the Measurement of Attitudes. NY: *Archives of Psychology*, *140*.

MacKenzie, S. B., Podsakoff, P. M., & Fetter, R. (1991). Organizational citizenship behavior and objective productivity as determinants of managerial evaluations of salespersons' performance. *Organizational Behavior and Human Decision Processes*, *50*, 123-150.

Mayfield, J., Mayfield, M., & Kopf, J. (1995). Motivating language: Exploring theory with scale development. *Journal of Business Communication*, *32*(4), 329-345.

Miles, M. B. & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd Ed.). Thousand Oaks: Sage Publications.

Miller, W. L. & Crabtree, B. F. (1999a). Using codes and code manuals: A template organizing style of interpretation. In B. F. Crabtree & W. L. Miller (Eds.), *Doing qualitative research* (2nd ed.), (pp. 163-178). Thousand Oaks, CA: Sage Publications.

Miller, W. L. & Crabtree, B. F. (1999b). Clinical research: A multimethod typology and qualitative roadmap. In B. F. Crabtree & W. L. Miller (Eds.), *Doing qualitative research* (2nd ed.), (pp. 3-30. Thousand Oaks, CA: Sage.

Minichiello, V., Aroni, R., Timewell, E., & Alexander, L. (1995). *In-depth interviewing* (2nd Ed.). Melbourne: Longman Cheshire.

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Patton, M. Q. (2002). *Qualitative research & evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Podsakoff, P. M., MacKenzie, S. B., Moorman, R. H., & Fetter, R. (1990). Transformational leader behaviors and their effects on followers' trust in leader, satisfaction, and organizational citizenship behaviors. *Leadership Quarterly*, *1*(2), 107-142.

Richards, L. & Richards, T. (1994). Using computers in qualitative analysis. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of Qualitative Research* (pp. 445-462). Thousand Oaks: Sage.

Schriesheim, C. A., Powers, K. J., Scandura, T. A., Gardiner, C. C., Lankau, M. J. (1993). Improving construct measurement in management research: Comments and a quantitative approach for assessing the theoretical content adequacy of paper-and-pencil survey-type instruments. *Journal of Management*, *19*, 385-417.

Schwab, D. P. (1980). Construct validity in organizational behavior. In L. L. Cummings & B. M. Staw (Eds.), *Research in organizational behavior*, Vol. 2 (pp. 3-43). Greenwich, CT: JAI Press.

Simon, A. M. (1990). A generative research strategy for data production. Illustrated by a Zimbabwean and a South African case study. In P. Hugo (Ed.), *Truth Be in the Field.* Pretoria, South Africa: UNISA.

Tabachnick, B. G. & Fidell, L. S. (1989). *Using multivariate statistics*. New York: Harper and Row.

Tinsley, H. A. & Tinsley, D. J. (1987). Uses of factor analysis in counseling psychology research. *Journal of Counseling Psychology*, *34*(4), 414-424.