A Methodology for Developing Normalized Formative Indices Using Messy Data

Completed Research

Gary F. Templeton
Mississippi State University
gftempleton@msstate.edu

Laurie Burney
Baylor University
Laurie_Burney@baylor.edu

Martin Kang
Mississippi State University
soporide@gmail.com

Abstract

The global expansion of information technology is providing MIS researchers greater accessibility to secondary data sources than ever before. Unfortunately, practically all of the variables in this era of Big Data are non-normally distributed and many have large proportions of missing values. In many cases, extreme distributional problems such as inflated frequencies (e.g., stacks of zeroes) inhibit statistical analyses for data mining or theory testing purposes. Because of the massive amounts of available 'messy' data, practitioners and researchers need methods for prioritizing and analyzing their archival datasets. Thus, we propose a normalized formative index development methodology that progresses through five stages: content specification, data collection, data reduction, technical validation and norming. Each stage of the proposed methodology is validated by its successful use in multidisciplinary research on formative index construction.

Keywords

Formative index, formative construct, normalize, structural equation modeling, methodology

Introduction

An investigation into MIS literature reveals serious impediments to scientific progress owing to formative measurement (Diamantopoulos, 2011; Kim, Shin & Grover, 2010), especially when using secondary data. Such impediments include missing data (Enders & Peugh, 2004), the lack of reflective measures useful in replication, and non-normality (Roy, 2008). While methods for formative index construction exist (Broby, 2007; Klumpner & Gedo, 1976), none have addressed the major issues plaguing archival data. To set forth best practices, we provide a comprehensive approach that calls for normality transformations applied to all index indicators to improve the reliability and predictive power of measures. We agree that "Further heuristics and guidelines for bringing even more rigor to the process of positivist, quantitative research need to be proffered" (Boudreau, Gefen & Straub, 2001, p. 13). Thus, the purpose of this paper is to articulate guidance on how to develop *normalized formative indices* using secondary data.[1] We urge researchers using secondary data to optimize construct validity and reliability in their indices and to standardize measures across studies through reuse. Such efforts can help research communities make better progress through generalization.

Indices are vital in theories across business disciplines where measures have been used to predict important organizational outcomes, such as firm failure rates (Wilcox, 1971) and investment

[1] Researchers not willing to transform their data to normality, for whatever reason, are referred to the large array of available non-parametric procedures. However, these procedures always involve distributional transformations that invariably transform data closer to the normal distribution ideal.

attractiveness (Palepu, 1986). Our approach is novel in that it is most useful where non-normality problems are severe. Use of this methodology may improve prediction and causal modeling compared to existing practices based on secondary data.

The paper describes the proposed methodology in order of logical progression that culminates with a validation of each part of the methodology with conclusions following. An illustrative example is provided as the methodology is described. The example progresses through the five proposed phases for normalized index construction applied to quarterly changes in the price/earnings (P/E) ratios of the thirty securities contained in the Dow Jones Industrial Average (DJIA). Thus, the example is an adaptation of perhaps the most prominent formative index used in practice and relates to the topic of information economics.

Methodology Progression

Five sequential phases will be described according to a prescribed sequence: I) content specification, II) data collection, III) data reduction, IV) validation, and V) norming. Adaptations are expected due to differences in underlying data (nominal, ordinal, interval, ratio), past research findings, and underlying distributions. Decisions early in the process strongly influence downstream results and decisions regarding the content and construct validities of the final index. In all phases, researchers should focus on enhancing the validity and reliability of the resulting index. For this reason, it is imperative for users of this approach to become familiar with its parts, to iteratively attempt each step, and to conduct further research if necessary.

Content Specification Phase

A formative index development process starts with content specification (Phase I), which has tremendous influence on content validity. According to (Agarwal & Venkatesh, 2002), content validity has "a powerful downstream influence on instrument validation" (Petter, Straub & Rai, 2007, p. 639). Content specification involves four steps aimed at establishing the conceptual scope of the theoretical construct measured by the index. Researchers begin by reviewing the literature (Diamantopoulos & Winklhofer, 2001; Petter, Straub & Rai, 2007) to establish a written domain definition (Step A) that theoretically encompasses the full range of meanings represented by the index. The definition should articulate what the index is as distinct from other concepts that may be in its nomological proximity.

Researchers then review the literature to compile the initial indicator pool (Step B). The goal is to include enough indicators so that as much of the index domain definition is represented as is practical. The capacity for the index to establish content validity (Straub, 1989) depends on the extent to which the index domain definition is represented by the initial item pool. Because indices should be constructed to reflect levels of the defined variable and not to influence findings (Broby, 2007), researchers should establish as much content validity as is practical.

Researchers then engage in domain testing (Step C), which is the evaluation of each item for representativeness within the domain definition. Using the Q-sorting procedure (Petter, Straub & Rai, 2007), items judged by domain experts as distinctly not representing the theoretical definition are cast out of the analysis. This step is an important way to ensure content validity before continuing to the technical analyses. The final element of content specification is establishing index dimensionality type and order (Step D). All formative indicators adhere to four decision rules: 1) each indicator causes the overall construct, 2) removing any indicator will remove meaning from the overall construct, 3) indicators may covary among one another, and 4) indicators may have different theoretical antecedents and consequences (Diamantopoulos & Winklhofer, 2001). The remainder of the methodology addresses the development of *2nd* order formative indices.

Applying the content specification phase to the example case reveals that each of the four steps is essential. To capture all of the contents of the resulting index (Step A), we named the target measure the *Normalized Dow Jones P/E Change Index (NDJPECI)*. We defined it as "a normalized stationary representation of the Dow 30." When defining a theoretical construct, it is important to understand its boundaries, including its position in a nomological network of interest to researchers. In finance, it has been theorized that P/E ratios can be used to predict the returns of individual stocks (DeBoeuf, Lee &

Stanley, 2013). Therefore, our example centered on developing an index of normalized P/E changes intended to predict overall market fluctuations. In this example, any economic antecedent, context, or consequence of P/E changes are expressly not a part of the domain of interest. Likewise, the construct of interest is differentiated from the quarterly returns of the Dow 30 Index. The initial indicator pool (Step B) was then established by compiling the historical quarterly P/E changes for each of the thirty stocks comprising the DJIA. A tenured business professor tested whether each item represented the domain definition (Step 3) using the Q-sorting technique. This resulted in the removal of non-representative items based on relevant evidence. An observation of descriptive statistics revealed that four indicators had unusually small sample sizes: Travelers (n=65), Cisco Systems (97), Visa (25), and Goldman Sachs (55). These items were removed from the analysis because the author made the subjective judgment that they were not representative.[2] In the context of the proposed methodology, the dimensionality (Step 4) of the *NDJPECI* is 2nd order of the formative type. The formative specification of the index is justified by each indicator being uniquely defined and determined as practical phenomena.

Data Collection Phase

Data collection (Phase II) involves three steps (A-C) that are informed by the initial content specification phase. Depending on circumstances, any of these steps may go a long way toward shaping the eventual index.

The extraction of all inputs (Step A) necessary for calculating the initial indicator pool is often complicated by the need for programming. Researchers first need to decide on the composition of subjects. Enough records should be collected to reach a sample size that allows acceptable statistical power, as well as records that are conceptually relevant (e.g., those in a geographical area or timeframe). One consideration is the need for balanced sample sizes across time and groups that will be used in validation. The composition of records should also consider assumptions of statistical tests used in analyses. For the index construction effort, researchers should also be concerned with the composition of necessary inputs, which depends on the formulae that define the initial indicator pool.

Deliberating among appropriate inputs may be a process of iterative refinement, as this step can be greatly complicated by four distributional phenomena unique to each input. First, inflated frequencies are recurring values within a variable that may not represent reality well. Second, missing values among inputs will have an unpredictable, but certainly detrimental, impact on the eventual sample sizes of calculated indicators. Third, lead and lag variables for each input should be calculated. Fourth, the presence of negative denominators can greatly distort the construct validity of indicators (Thornblad, Zeitzmann, & Carlson, 2018).

We suggest that researchers inspect the causes of extensive inflated frequencies. Where these values are illogical, researchers may consider deleting them. However, these values may also reflect real phenomenon and be essential to study results and interpretations. To address the daunting problem of missing data, we encourage single imputation utilizing multiple regression (Baraldi & Enders, 2010; Enders, 2010). That is, researchers are encouraged to use multiple regression to estimate missing values by leveraging information in higher sample sized inputs in the dataset (Gelman & Hill, 2006). Concerning leads and lags, researchers should draw realistic expectations about what timeframe will be useful in assessments. The exact time periods used will depend on theoretical or logical considerations. This step is may be done at the input or indicator (post-calculation) level, depending on the research context. Finally, we advocate one rule for addressing negative denominators: *transform any denominator to a uniform probability before a division is executed*. We suggest this rule for any denominator during ratio calculations, including those involve multiple division operations. This has the critical effect of retaining the logical order of ratios during calculation.

Step B of the Data Collection Phase of index construction involves establishing the initial indicator pool. The input variables are used to calculate the indicators as well as the exogenous variables employed for validation. An important decision made in constructing indices in a time-series context is whether the measure should represent stationary or non-stationary phenomena (Momin & Chavan, 2018). Stationary data is constructed by calculating changes from previous time periods and results in a distribution with

[2] Items could also have been removed for other reasons, such as operations (service vs. manufacturing), industry sector, or empirical distinctions (e.g., size).

well defined (i.e., stationary) boundaries. Alternatively, non-stationary measures may increase with no boundary and are usually based on original indicator formulae. Formative indices should not mix these types, as they are considered two distinct worldviews of measurement.

Step C of data collection involves assessing and addressing states of non-normality among calculated indicators. In some disciplines (e.g., finance), archival datasets contain levels of non-normality that are extraordinary by scientific standards (Barnes, 1982; Deakin, 1976). Yet, normality is an important assumption of parametric statistical procedures across all disciplines. To address normality, even in the most extreme circumstances, Templeton (2011) advocates a Two-Step transformation. This simple approach retains original unit means and standard deviations, is applicable to non-positive values, and is accessible among the popular statistical tools (e.g., Excel and SPSS). Using this procedure, researchers are provided two forms of units: 1) original and 2) standardized (μ=0, σ=1). This methodology utilizes the standardization option. Using standard-normal scoring has the advantage of securing the consistency of probability distributions across the index indicators.

Typical normality tests include those commonly available for skewness (which assess distributional asymmetry) and kurtosis (for thickness of distributional tails), and the omnibus Kolmogorov–Smirnov distributional test. Assessing normality both before and after transformation will illuminate the extent to which the procedure affected the distribution as well as the resulting status of normality among measures.

Inputs for the *NDJPECI* project were extracted (Step A) from the *Wharton Research Data Services (WRDS)*[3]. In order to capture the largest sample possible, the records span from the first quarter of 1961 to the second quarter of 2014. Indicators were calculated (Step B) using accepted understandings in business practice. For each stock, the P/E ratio was calculated as follows:

$$P/E = \frac{\text{Price at Quarter End}}{\text{Basic Earnings Excluding Extraordinary Items (12 Months Moving)}}$$

To arrive at the test values, each P/E ratio was calculated as a change relative to the previous quarter:

$$P/E\Delta = \frac{\text{Pri(P/E in current quarter} - \text{ P/E in previous quarter)ce at Quarter End}}{\text{P/E in Previous Quarter}}$$

The indicator pool was considered relatively clean, having no prominent inflated frequencies and the project did not involve lead or lag calculations. Missing values, however posed a problem for four indicators. Rather than engage in an arduous process of ameliorating missing values via the aforementioned simple imputation using multiple regression or mean replacement (which would suppress effect sizes), the decision was made to discard four indicators. This decision was acceptable due to the availability of indicators representing similar meaning in this context. To address any non-normality issues (Step C), the skewness and kurtosis of each remaining NDJPECI indicator was assessed. Due to excessive non-normality among all indicators, each were transformed using the Two-Step normalization procedure. The normality diagnostics for original and transformed versions revealed a vast improvement in normality. Among all of the 52 skewness and kurtosis diagnostic tests applied, the transformed distributions were normally distributed.

Data Reduction Phase

Phase III involves reducing the initial indicator pool to a parsimonious subset of representative surrogates that will comprise the index. Before data reduction, the problem is often a lack of direction as to which indicators to use to measure the overall construct. Each formative indicator has unique meaning, so removing any one will also remove some aspect of the index.[4] However, representing every conceivable part of any formative index is often impractical or impossible. A well-conceived process of data reduction can accomplish six objectives of interest to researchers: 1) advance prior research on the index by establishing its dimensions, 2) ensure that the surrogate indicators represent diverse components of the index (content validity), 3) achieve as much parsimony in the index as possible, 4) confirm that each

[3] WRDS is available on a subscription basis at https://wrds-web.wharton.upenn.edu/wrds/

[4] An important goal is to remove as much shared meaning (variance) among the indicators while also retaining as much unique meaning.

indicator is representative of a unique component of the index, 5) evaluate construct validity, including reliability, and 6) optimize the number of listwise records available when using the index and hence, the sample size-to-items ratio that will improve statistical power in subsequent research.

Five steps are involved in data reduction for index construction. First, **w**e use *principal components analysis (PCA)* to arrive at components among indicators (Step A).[5] We use the orthogonal varimax rotation for three reasons: 1) it is recommended for the development of formative indices (Petter, Straub & Rai, 2007), 2) it has widespread use in the social sciences, and 3) it produces uncorrelated components. Contrary to the iterative process usually employed to validate reflective scales, we conduct a single PCA analysis using all indicators in the initial pool. Second, researchers should investigate several conceptual and statistical assumptions associated with PCA analysis (Step B). Regarding conceptual assumptions, the PCA structure should be stable with clear components and few cross loadings. Regarding statistical assumptions, standard tests are performed to determine the extent to which the dataset is appropriate for applying PCA - the Kaiser-Meyer-Olkin and Bartlett sphericity tests. The sample size should be at least ten times larger than the number of variables (Hair, Black, Babin, & Anderson, 2010). Third, researchers should evaluate five criteria from the PCA output: component weights, eigenvalues, the scree test, percentage of total variance explained (PTVE), and alternative solutions (Step C). For each component, a single surrogate is proposed based on representativeness (using component weights), uniqueness (using variance inflation factors), and impact on listwise sample size (using indicator 'inclusion effect' on sample size). The group of surrogates is considered representative of the target index to the extent that the initial indicator pool is complete. Fourth, the interpretation and labeling of each component can be done both conceptually and empirically (Step D). From a conceptual standpoint, these components are represented by one index surrogate. To provide a label for each component, researchers may distribute the list of indicator groupings (by component) to experts in the field. Using an expert panel in this way also serves as a follow up to the content validity test (Petter, Straub & Rai, 2007). Fifth, researchers will choose one surrogate to represent each component (Step E). This does not discard information, as any remaining indicators serve as alternate representatives of the component in subsequent studies. Chosen surrogates should have a high absolute component weight (Hair, Black, Babin, & Anderson, 2010, p. 103) and large sample size. The surrogate set should explain at least 50% of the variance in the index it represents (Wierenga & van der Lans, 2017). We suggest using the Kullback-Leibler Divergence (KLD) to compare each surrogate to its parent PCA component to assess their degree of sameness (Hayashi, Bentler, & Yuan, 2007; Akaike, 1987). A lower KLD represents similarity between the surrogate and the composite and the extent to which the data reduction process retained original information. Being careful to maximize the sample size mitigates downstream sample issues during empirical testing.

Principal Components Analysis (PCA) using varimax (orthogonal) rotation in SPSS™ was applied to the normalized versions of the remaining 26 P/E return indicators (Step A). The assessment of conceptual and statistical assumptions (Step B) revealed a high degree of conceptual association due to their use in the DJIA composite. Empirically, the KMO value for this test was 0.670, indicating that the normalized dataset was amenable to the PCA procedure. Based on the preponderance of evidence, the PCA produced nine factors explaining 69.9% of total variance (Step C). Arriving at appropriate labels and definitions for the factors (Step D) is a matter of industry experience with firms in the DJIA composite. In the case of the NDJPECI, each was labeled with a number since no discernable theme was found among most factors[6]. Within each factor, the selection of surrogates among alternatives (Step E) was straightforward. The most explanatory (highest loading) indicator in each factor served as the surrogate for each factor.[7]

Validation Phase

Phase IV of formative index development involves validity testing to ensure that the index represents the target concept of meaning and behaves as theoretically intended.

---

[5] Guidance pertaining to the extraction method is (Hair, Black, Babin, and Anderson, 2010, p. 117): "Component analysis is used when the objective is to summarize most of the original information (variance) in a minimum number of factors for prediction purposes."

[6] The exception was the 4th factor, comprised of Chevron and Exxon in the oil and gasoline industry:.

[7] This clean approach was likely due to the exclusion of low-sample size indicators during Domain Testing (Content Specification Phase, Step C).

Content validity (Step A) is a critical first consideration in validity testing (Agarwal & Venkatesh, 2002). The indicators in the initial pool are explicitly evaluated for content validity in three earlier methodology phases: domain testing (Phase I, Step C), the PCA procedure (Phase III, Steps A-C), and component labeling (Phase III, Step D). These face validity checks offer external reviewers multiple opportunities to provide feedback on each indicator's fit with the overall index.

These influences are apparent in the development of the NDJPECI. Based on the initial inclusion of all securities in the DJIA, we know that any group of surrogates for P/E change will fit within the established domain definition. Transformations of all indicators were done using original series means and standard deviations to preserve the interpretation of empirical results in original units of analysis. Finally, we know that using seven items to represent the DJIA P/E Change Index demonstrates greater content validity than using fewer components. While reducing the size of the index from 26 to 7 items somewhat reduces content validity, the mission of this index development project was to provide for computational efficiency.

Construct validity (Step B) involves empirical tests to evaluate the extent to which the measure is internally and externally valid. A useful formative index will pass a plurality of unique checks aimed at assessing whether it behaves as intended. Researchers will assess factors internal and external to the index. Statistical conclusion validity is an internal consideration of the extent to which an index is useful in making correct conclusions (Shadish, Cook, & Campbell, 2002). Two characteristics of indices strongly influence statistical conclusion validity in studies. First, the index should not contribute to the violation of parametric assumptions. This is chiefly addressed in the methodology via normality transformation. Extremely non-normal data such as that commonly found in the economic domain negatively influences both covariance- and component-based structural equation modeling (Qureshi & Compeau, 2009). Second, statistical power is the conditional probability of making the correct finding in a hypothesis test (Cohen, 1988). Low statistical power is directly related to the available sample size (Aberson, 2011; Baroudi & Orlikowski, 1989) and consequently, missing values. Statistical power is enhanced by the estimation stability of formative measures (Kim, Shin & Grover, 2010). Therefore, researchers should also assess autocorrelation among surrogates used in analyses.

Three primary types of external construct validity relate to formative indices. First, convergent validity is the extent to which multiple methods of measuring the same concept agree using empirical tests (Straub, Boudreau & Gefen, 2004). A component weight in PCA can be interpreted as the level of agreement between the indicator and the dimension of the index (Straub, Boudreau & Gefen, 2004). Thus, two or more indicators having significant weights on a component are showing convergent validity. Second, discriminate validity is the extent to which an indicator differs from other indicators as it should. Establishing this validity is largely accomplished in the data reduction phase by using the orthogonal rotation method. As long as the surrogates do not cross-load, the researcher has established evidence of discriminate validity (Straub, Boudreau & Gefen, 2004). Third, nomological validity is the extent to which we can confirm that the index is useful in the prediction of theoretical consequences. Nomological validity can be assessed using traditional multiple regression and path analysis approaches modeling surrogates as independent variables and theoretical outcomes as the dependent variables. Treiblmaier, Bentler & Mair, 2011) have described an acceptable approach for estimating the measurement models of formative indices using covariance-based structural equation modeling. While this validation approach fits the formative measurement philosophy very well, a simpler option is to compute a summary score for the index and model it as a single variable in the theoretical model.

The NDJPECI measure was assessed for internal and external construct validity. None of the indicators demonstrated violations of the normality assumption. The methodology also designed the NDJPECI to address the varying amounts of missing values its users will encounter that will impair working sample sizes. This was accomplished in part by the PCA procedure, which dramatically reduced the number of working indicators. This increases statistical power as well as the sample-to-item ratio in subsequent research. These empirical accomplishments all positively affect the construct validity of NDJPECI measures. Regarding external construct validity, a measurement model for the NDJPECI was estimated. We relied on exploratory factor analyses (EFA) using SPSS™ 22.0 to split the variables in a way to minimize inter-correlation within the composites, while maximizing it between composites. Then, we generated a MANOVA to estimate the canonical correlation. Once determined, a more specific structural model was developed for estimating. The structural equation model resulted in a model with a squared multiple correlation for the Dow Index Quarterly Return of 91.9%. The model provided a mediocre fit to the data with an acceptable CMIN/DF of 3.987 and a high root mean square error of approximation (RMSEA) value of 0.117[8].

Reliable formative indices demonstrate internal and external construct validity consistently across groups and over time (Step C). If results are consistent over groups and time, validity is stronger because it has been replicated over different parts of the sample. Conversely, if results over groups or time periods are sporadic, evidence indicates poor validity. Each of these types of validity can be evaluated at the indicator level, but it can also be demonstrated at the dimension (component) and construct (index) levels (Straub, Boudreau & Gefen, 2004). Since the formative model is based on regression (Bollen & Lennox, 1991), an important goal of formative index development is to produce indicators with very little multicollinearity. Each indicator in the surrogate subset is chosen to be as independent as possible during data reduction, as ultimately demonstrated in external validity tests. This is accomplished in large part by using PCA and the orthogonal varimax rotation approach. Researchers may initially assess the index's reliability during the data reduction stage by assessing the stability of the PCA structure in two basic ways. First, researchers can calculate the difference between the total percentage of variance explained (TPVE) in the final model and the same figure for random variates. Generally, the larger this difference, the more variance the components explain; and the indicators will predict the overall construct better. Second, researchers can remove problematic indicators from the PCA analysis and observe changes (in weights and component composition) to the remainder of the model. If a reduced model remains essentially the same as the larger model, the component structure is consistent. When later used in practice, this may be assessed using variance inflation factors (VIFs).

Another stabilizing aspect of indices created using the above approach is normalization. The consequences of this will be shown as improved consistency in statistical tests. To explicitly address reliability, researchers may use the index to assess test-retest reliability[9], which investigates the stability of the formative index over time or across groups. To assess the stability of a component model, the estimation results are compared across a number of subsets from the study sample data. Given that the individual indicators capture a different aspect of the overall index, a comparison at the surrogate level is preferred over the index level for both types of test-retest reliability (Petter, Straub & Rai, 2007). The overall results are stable if they are consistent over time or among samples.

The NDJPECI surrogates were regressed against the DJIA returns for the concurrent year to obtain variance inflation factors (VIFs) for each predictor. VIFs ranged from 1.06 to 1.17, which indicated that the beta coefficients in the regression model should be considered reliable. The 26 indicators for stock returns were entered into a principal component analysis (PCA) using varimax rotation. This procedure was performed twice, once each for original and normalized versions of variables. Using the original data as a baseline, the normalized factorial structures were more stable when individual items were removed. This suggested stability in the factorial structures. The dataset was segmented into four groups, one for each quarter (1, 2, 3 and 4) of a calendar year. We applied independent samples differences tests to assess the extent of the variance and means equality between quarters for each surrogate. Four sets of tests were conducted: 1) between Quarter 1 (Q1) and Q2, 2) Q2 and Q3, 3) Q3 and Q4, and Q4 and Q1. The F statistic for equality of variances (column heading "σ^2 *F*") was insignificant in 34 of 36 (94%) Levene's tests. Further, the t-tests for means (column heading is "μ *t*") were insignificant in 32 of 36 (88%) tests. These results show that the measures are temporally consistent between quarters for both variances and means.

Norming Phase

Finally, the need to calculate norming data when organizational indices are developed (Lewis, Templeton & Byrd, 2005) is addressed. For example, finance and accounting literature has found that financial ratios are expected to change over time (Tippett, 1990) and are sensitive to industry conditions (Wu & Ho, 1997). Thus, time and competitive context are important factors to consider when utilizing the index in

---

[8] CMIN/DF should be 5 or lower (Wheaton, Muthén, Alwin & Summers, 1977). RMSEA should be lower than 0.07 (Steiger, 2007).

[9] The test-retest reliability of formative indices is important in MIS because there have traditionally been so few longitudinal studies (Pinsonneault & Kraemer, 1993).

studies. For both untransformed and transformed values, the means and standard deviations should be estimated and reported for each of the index surrogates. These descriptive statistics should be calculated for all records and may be grouped by time or other criterion (e.g., industry).

The NDJPECI dataset was again segmented into four quarterly groups. Means and standard deviations were calculated for each quarter. Quarterly norms for means, standard deviations, and sample sizes across these groups were calculated. These values can be used as company norms for comparison.

Methodology Assessment

The proposed methodology is justified through a combination of approaches found in the research methodology literature. While not used collectively, each methodological phase, step and procedure has been used to accomplish the same objectives in published research. The methodology is heavily supported by two-staged development methodologies intended for psychometric data, focusing on: 1) formative scales (Diamantopoulos & Winklhofer, 2001) and 2) reflective scales (Lewis, Templeton & Byrd, 2005). While these guides remain useful in their own right, our methodology has the specific purpose of developing normalized formative scales using archival (highly continuous) data. As such, we believe it is applicable to most of the data available to business practitioners as the Big Data era expands. Similar approaches have been used in the relatively short history of developing and validating formative indices in the MIS literature (Agarwal & Venkatesh, 2002; Bock, Zmud, Kim & Lee, 2005; Pavlou & Fygenson, 2006).

Specific procedures that we advocate are also supported in previously published research on research methodology. The Data Reduction phase was heavily inspired by (Hair, Black, Babin, & Anderson, 2010) and advocated in prior literature (Petter, Straub & Rai, 2007). The Two-Step transformation toward normality set forth by Templeton (2011) is essential for developing indices based on continuous data that are reliable and externally valid.

Discussion and Conclusions

Our goal is to articulate proposed steps, procedures and associated logic for developing normalized formative indices. The proposed methodology facilitates researcher control over the historical problems of using archival data, such as small sample size, missing data, and non-normality. The methodology guides researchers to empirically prioritize and select specific indicators among alternatives capturing a domain of interest. Among subsequent studies, researchers should adhere to the proposed subset of surrogates to better compare results and build a cumulative tradition of econometrics research. Among the many examples of archival data characterized by non-normality are data collected from remote sensors in agriculture (e.g., https://www.data.gov/food/), atmospheric sciences (see the National Oceanographic Data Center: http://www.nodc.noaa.gov/), healthcare (see http://www.hcup-us.ahrq.gov/databases.jsp), and transportation (see TranStats: http://www.transtats.bts.gov/). In business, most of the data in the Wharton Research Data Services databases (e.g., COMPUSTAT, CRSP, and Eventus) are extremely non-normally distributed, greatly affecting findings of research streams.

This methodology is a basic guide for practically any topic using secondary data since statistical normality[10] is rare among archival data in some disciplines. A good example is corporate financial performance, which is a domain known to rely on indicators having extremely non-normal distributions. Our approach is especially invaluable since to date, no practical situations are known where archival variables have known interdependencies when all of the indicators have been transformed toward normality. Researchers may also adapt the proposed methodology to projects using primary data by altering traditional perceptual scales to have 100 or 1,000 levels instead of the usual five or seven. After data collection, the items would then be transformed toward normality and developed into a normalized formative index. Researchers may opt to use a single surrogate to represent each previously validated

[10] Distributions in many published IS studies are often judged as normal without strictly adhering to strict diagnostic procedures (e.g., kurtosis, skewness, and the Kolmogorov-Smirnoff normality test). Therefore, we consider 'statistical normality' to be a strict standard.

reflective construct in a multidimensional superordinate construct as in the IT Relatedness scale (Tanriverdi, 2006).

Although the design and refinement of normalized formative indices can be arduous, the outcome may have tremendous benefits. First, the development of formative indices requires researchers to conceptually define and validate the measure, which promises to vastly improve measure reliability. Second, developed formative indices mitigate information overload by summarizing or reducing the amount of information consumed. By proposing a justifiable subset of the data available, comparisons are allowable across organizational units or scientific studies. Third, successful formative index development enables researchers to establish norms for different nationalities, industries and other organizational demographic variables.

Due to the wide range of situations encountered with secondary data, the methodology has its limitations. First, using PCA for the critical step of data reduction does not show the researcher which of the derived components or surrogates are predictors and which are criterion. Therefore, the findings of this research have limited usefulness in research contexts that involve specific relationships between index components (surrogates). Such would be the subject of future research for each index developed using the methodology. Second, studies focused on a specific industry or time period may not find an index developed for general purposes as useful. For instance, a general-purpose IS investment index developed using this methodology may not have the same success in a study within the healthcare industry. Further, an index created using this approach may need to be updated due to changing dynamics in new economic climates. Finally, researchers should caution against using a single composite measure to represent an entire index. This approach will prevent researchers from making inferences about the distinct index dimensions. We hope future uses of the methodology will spawn new discoveries and fuel subsequent research that will improve formative index construction using continuous data.

REFERENCES

Aberson, C.L. 2011. *Applied power analysis for the behavioral sciences,* Routledge.

Agarwal, R. and Venkatesh, V., 2002. "Assessing a firm's web presence: a heuristic evaluation procedure for the measurement of usability," *Information Systems Research*, (13:2), pp. 168-186.

Akaike, H. 1987. *Factor analysis and AIC*, in *selected papers of Hirotugu Akaike*, Springer, New York, NY.

Baraldi, A.N. and Enders, C.K. 2010. "An introduction to modern missing data analyses," *Journal of school psychology* (48:1), pp. 5-37.

Barnes, P. 1982. "Methodological implications of non-normally distributed financial ratios," *Journal of Business Finance & Accounting* (9:2), pp. 51–62.

Baroudi, J. J. and Orlikowski, W. J. 1989. "The Problem of Statistical Power in MIS Research," *MIS Quarterly* (13:1), pp. 87–106.

Bock, G.-W., Zmud, R. W., Kim, Y.-G., and Lee, J.-N. 2005. "Behavioral intention formation in knowledge sharing: Examining the roles of extrinsic motivators, social-psychological forces, and organizational climate," *MIS Quarterly* (29:1), pp. 87–111.

Bollen, K. and Lennox, R. 1991. "Conventional wisdom on measurement: A structural equation perspective," *Psychological Bulletin* (110:2), pp. 305–314.

Boudreau, M.-C., Gefen, D., and Straub, D.W. 2001. Validation in Information Systems Research: A State-Of-The-Art Assessment. *MIS Quarterly* 25(1), pp. 1-16.

Broby, D. 2007. *A Guide to Equity Index Construction*, London: Risk Books.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd *Ed.* Hillsale, NJ: Lawrence Erlbaum.

Deakin, E. 1976. "Distributions of financial accounting ratios: Some empirical evidence," *The Accounting Review* (51:1), pp. 90–96.

Diamantopoulos, A. 2011. "Incorporating Formative Measures into Covariance-Based Structural Equation Models," *MIS Quarterly* (35:2), pp. 335-A5.

Diamantopoulos, A. and Winklhofer, H. M. 2001. "Index Construction with Formative Indicators: An Alternative to Scale Development," *Journal of Marketing Research* (38:2), pp. 269–277.

Enders, C. K. and Peugh, J. L. 2004. "Using an EM covariance matrix to estimate structural equation models with missing data: Choosing an adjusted sample size to improve the accuracy of inferences," *Structural Equation Modeling* (11:1), pp. 1–19.
Enders, C. K. 2010. *Applied Missing Data Analysis,* New York: Guilford Press.
Gelman, A. and Hill, J. 2006. *Data analysis using regression and multilevel/hierarchical models,* Cambridge University Press: Cambridge.
Hair, J.F., Black, W.C., Babin, B.J., and Anderson, R.E. 2010. *Multivariate Data Analysis,* 7th Ed. Upper Saddle River, NJ: Prentice Hall.
Hayashi, K., Bentler, P. M., and Yuan, K. H. 2007. "On the likelihood ratio test for the number of factors in exploratory factor analysis," *Structural Equation Modeling: A Multidisciplinary Journal* (14:3), pp. 505-526.
Kim, G., Shin, B., and Grover, V. 2010. "Investigating Two Contradictory Views of Formative Measurement in Information Systems Research," *MIS Quarterly* (34:2), pp. 345-A5.
Klumpner, G. H. and Gedo, J. E. 1976. "Toward a conceptual index of psychoanalytic terms: A method to develop a psychoanalytic thesaurus," *Journal of the American Psychoanalytic Association*, (24:2), pp. 409–424.
Lewis, B. R., Templeton, G. F., and Byrd, T. A. 2005. "A methodology for construct development in MIS research," *European Journal of Information Systems* (14:4), pp. 388–400.
Momin, B. and Chavan, G. 2017, "Univariate Time Series Models for Forecasting Stationary and Non-stationary Data: A Brief Review," in *International Conference on Information and Communication Technology for Intelligent Systems*, pp. 219-226.
Palepu, K. G. 1986. "Predicting Takeover Targets: A Methodological and Empirical Analysis," *Journal of Accounting and Economics* (8:1), pp. 3–35.
Pavlou, P. A. and Fygenson, M. 2006. "Understanding and Prediction Electronic Commerce Adoption: An Extension of the Theory of Planned Behavior," *MIS Quarterly* (30:1), pp. 115–143.
Petter, S., Straub, D., and Rai, A. 2007. "Specifying Formative Constructs in Information Systems Research," *MIS Quarterly* (31:4), pp. 623–656.
Pinsonneault, A. and Kraemer, K. L. 1993. "Survey Research Methodology in Management Information Systems: An Assessment," *Journal of Management Information Systems* (10:2), pp. 75–105.
Qureshi, I. and Compeau, D. 2009. "Assessing Between-Group Differences in Information Systems Research: A Comparison of Covariance- and Component-Based SEM," *MIS Quarterly* (33:1), pp. 197–214.
Roy, S. 2008. "Measuring Formative Constructs in Management Research: Definitions, Distinctions and Measurement," *ICFAI Journal of Management Research* (7:10), pp. 18–34.
Shadish, W., Cook, T.D., and Campbell, D.T. 2002. Experimental and quasi-experimental designs for generalized causal inference. *Boston*.
Straub, D. 1989. "Validating Instruments in MIS Research," *MIS Quarterly* (13:2), pp. 147–169.
Straub, D., Boudreau, M.-C., and Gefen, D. 2004. "Validation Guidelines for IS Positivist Research," *Communications of the Association for Information Systems* (13:1), pp. 380–427.
Tanriverdi, H. 2006. "Performance Effects of Information Technology Synergies in Multibusiness Firms," *MIS Quarterly* (30:1), pp. 57–77.
Templeton, G. F. 2011. "A Two-Step Approach for Transforming Continuous Variables to Normal: Implications and Recommendations for IS Research," *Communications of the Association for Information Systems* (28:1), pp. 41–58.
Thornblad, D. B., Zeitzmann, H. K., and Carlson, K. D. 2018. "Negative Denominators in Index Variables: The Vulnerability of Return on Equity, Debt to Equity, and Other Ratios," *The Electronic Journal of Business Research Methods* (16:1), pp. 1-10.
Tippett, M. 1990. "An Induced Theory of Financial Ratios," *Accounting & Business Research* (21:81), pp. 77–85.
Treiblmaier, H., Bentler, P. M., and Mair, P. 2011. "Formative constructs implemented via common factors," *Structural Equation Modeling* (18:1), pp. 1–17.
Wierenga, B. and van der Lans, R. 2017. Handbook of Marketing Decision Models. Springer.
Wilcox, J. W. 1971. "A Simple Theory of Financial Ratios as Predictors of Failure," *Journal of Accounting Research* (9:2), pp. 389–395.
Wu, C. and Ho, S.-J. K. 1997. "Financial Ratio Adjustment: Industry-Wide Effects or Strategic Management," *Review of Quantitative Finance and Accounting* (9:1), pp. 71–88.