Archives of Business Research – Vol.6, No.10
Publication Date: Oct. 25, 2018
DOI: 10.14738/abr.610.5319.
Hensley, R. L., & Utley, J. S. (2018). Assessing Multiple-Item Manufacturing Capability Measures: A Critical Review. *Archives of Business Research, 6(10),* 202-227.

Assessing Multiple-Item Manufacturing Capability Measures: A Critical Review

Rhonda L. Hensley
Associate Professor, Department of Management,
College of Business and Economics, North Carolina A&T State University

Joanne S. Utley
Professor Emeritus, Department of Management,
College of Business and Economics, North Carolina A&T State University

ABSTRACT

This paper critically reviewed empirical studies of manufacturing capabilities to identify possible inconsistencies in the composition and use of multiple item capability measures. In addition, this paper examined the issue of item commonality which is the extent to which questionnaire items corresponding to a capability were used by the studies to measure the capability. This paper also explored the potential effects of inconsistencies in measure composition and use on survey validity and on item commonality. Twenty-seven papers met the review criteria delineated in the paper. Review results revealed a striking lack of consensus in measure sourcing, Likert scale framing and the use of validity checks in measure composition. Furthermore, only six survey items used in the studies can be thought of as high commonality items. Given these results, the review offers recommendations for future research on capability measurement.

Keywords: Capability measures; Empirical research; Measurement development; Manufacturing capabilities; Item commonality

INTRODUCTION

It has long been recognized that manufacturing capabilities play a key role in the performance of manufacturing firms (Skinner, 1969; Anderson et al., 1989; Roth and Miller, 1992; Vickery et al., 1993; O'Regan and Ghobadian, 2004; Ibrahim, 2010; Bronzo et al., 2012). Manufacturing capabilities are "the strengths of a plant with which it wants to support corporate and marketing strategy and which help it to succeed in the marketplace" (Größler and Grübner, 2006, p. 459). Manufacturing strategy research has identified quality, delivery, flexibility and cost as the most widely accepted capabilities and has sought reliable methods to measure manufacturing capabilities and to compare them in actual business contexts (Größler and Grübner, 2006).

The ability to reliably measure manufacturing capabilities remains important for both practical and theoretical reasons. Effective capability measurement can help a firm make sound decisions regarding capacity, process choice, technology, quality and manufacturing planning and control systems (Skinner, 1969; Hayes and Wheelwright, 1984; Ward et al., 1998). Theoretically, sound capability measures are essential for analyzing and "explaining patterns of capability development" in manufacturing firms (Amoako-Gyampah and Meredith, 2007, p. 929) and for testing and comparing competing theoretical models of capability development.

Given their importance, how are manufacturing capabilities best measured? In recent years researchers have moved away from single-item capability measures and toward multiple-item measures. Schroeder et al. (2011, p. 4888-4889) offer a rationale for this trend by noting

"using single indicators to operationalize manufacturing performance is restrictive because it does not adequately capture the breadth of each manufacturing performance dimension."

Despite the importance of manufacturing capabilities, the literature lacks critical reviews that focus squarely on capability measure development and use in empirical research. This is surprising since previous research has suggested that capability measurement can be a flawed process. For instance, in their review of studies testing capability models, Sarmiento et al. (2010, p. 1282) found "a lack of appropriate operationalization of scales and variables" for measuring capabilities. Similarly, a meta analysis of studies of manufacturing performance and tradeoffs by Rosenzweig and Easton (2010, p. 136) reported that "researchers often utilize different variable operationalizations for multidimensional complex constructs." In addition, Sarmiento et al. (2010) called for more consistent terminology, the use of external frames of reference in assessing capability and longitudinal and case study approaches. Boyer and Pagell, (2000) raised concerns about the predictive validity of capability measures as well as inconsistencies in constructing surveys of manufacturing capability.

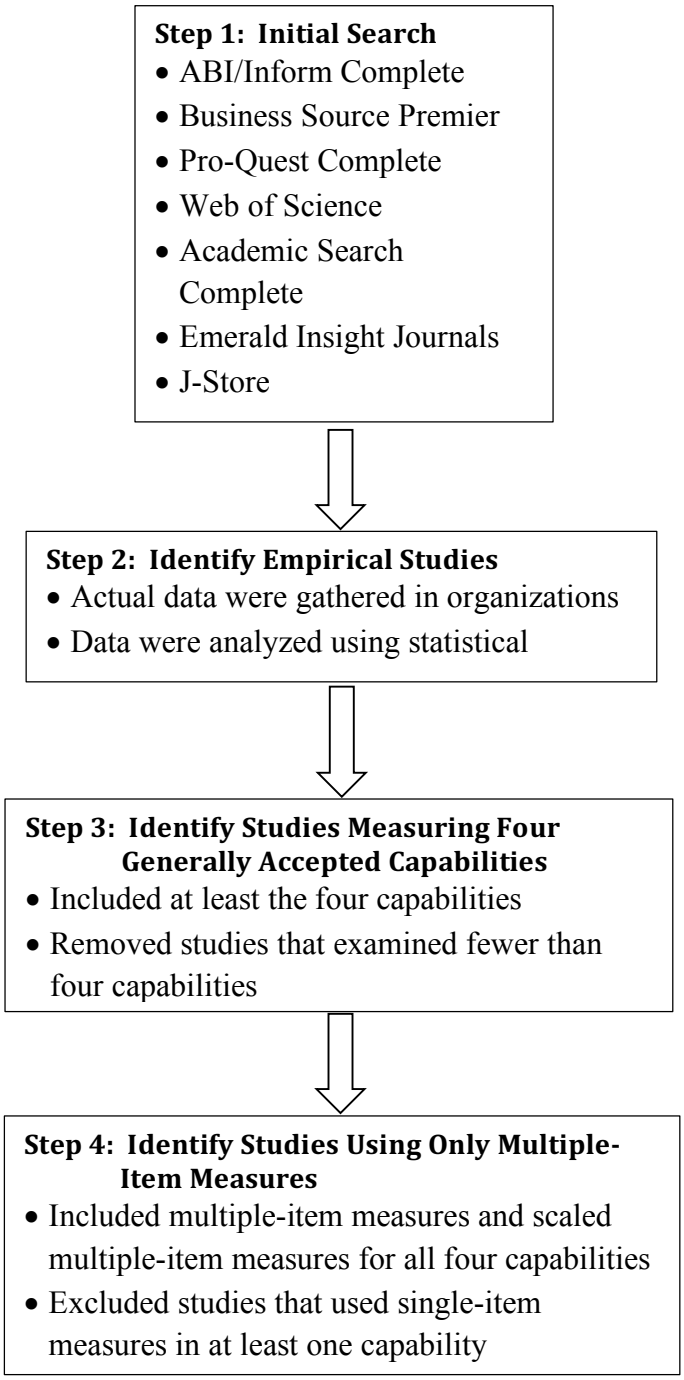
The potential flaws in capability measurement described above give rise to two important lines of inquiry. First, to what extent are inconsistencies in measure composition and use found in empirical studies that use multi-item capability measures? Second, how much commonality is there among the items used in the measures in these studies? This paper will critically review empirical studies that used multiple-item capability measures to address both of these questions. The paper will also examine the possible effects of any inconsistencies in measure composition and use that are found; furthermore possible reasons for any commonality differences among the items used in the measures will be explored.

The remainder of this paper is organized as follows. The following section describes the methodology used to select the studies for review. Section three examines the composition and use of capability measures in the review papers. Section four presents an assessment of item commonality of the measures in the studies under review. Section five examines the possible ramifications of inconsistencies in measure composition and use in the review studies as well as the reasons for variations in item commonalities for the measures in these papers. Section six provides a discussion of the review findings. The paper concludes with recommendations for future research on manufacturing capability measurement.

METHODOLOGY FOR IDENTIFICATION OF RELEVANT PAPERS

A four step process was followed to identify studies measuring manufacturing capabilities (Figure 1). First, eight online databases were searched, ABI/Inform Complete, Business Source Premier, Pro-Quest Complete, Web of Science, Academic Search Complete, Emerald Insight Journals, J-Stor and Science Direct. A separate search of operations management journals was conducted to ensure thoroughness in identifying articles. The references contained in the papers found by these two searches were further scrutinized to identify any other papers that should be included in the critical review.

Figure 1
Selection and Review of Studies



In the second step, those papers defined as empirical were identified and retained. The importance of empirical studies in the development of operations management theory has been well documented in the literature (see, for instance, Flynn et al., 1990; Schmenner and Swink, 1998; Wacker, 1998; Meredith, 1998; Fisher, 2007; Meredith, 2009; Boer et al., 2015). Minor et al. (1994) defined empirical studies as those studies in which actual data were gathered from real world organizations and analyzed using statistical techniques. Using this definition, over 60 studies were identified.

In the third step, studies that failed to measure all four generally accepted manufacturing capabilities were eliminated. Studies that either measured the four capabilities or measured the four but included additional capabilities were retained.

Finally, in step four, studies using single-item measures were eliminated from further review. Since this paper examines the different item measures for the four capabilities, it was important to examine only those studies using multiple-item measures. The final list of studies included 27 studies measuring all four capabilities with either multiple-item measures or scaled multiple-item measures. This final list appears in Table 1.

Table 1
Sources of Measures

Developed Measures for Particular Study	Adopted Measures from Previous Studies
Ward et al. (1995)	Noble (1995)
Ward et al. (1998)	Noble (1997)
Katuria (2000)	Boyer (1998)
Sum et al. (2004)	Boyer & McDermott (1999)
Größler & Grübner (2006)	Corbett & Whybark (2001)
Amoako-Gyampah & Meredith (2007)	Boyer & Lewis (2002)
Swink et al. (2007)	Rosenzweig et al. (2003)
Chung & Swink (2009)	Squire et al. (2006)
Avella et al. (2011)	Avella & Vazquez-Bustelo (2010)
Schroeder et al. (2011)	Liu et al. (2011)
Sum et al. (2012)	Wong et al. (2011)
Wu et al. (2012)	Singh et al. (2014)
Chavez et al. (2017)	Boon-itt & Wong (2016)
	Ehie & Muogboh (2016)

COMPOSITION AND USE OF CAPABILITY MEASURES

A useful starting point in examining capability measurement involves answering the question: where did the capability measures in the review papers come from? Table 1 shows that the measures came from two different sources. Thirteen papers developed capability measures as part of their study. In these papers, the questionnaire items used in each capability measure were tailored to the individual study. In contrast, fourteen studies used measures from either pre-existing databases or from earlier studies.

In addition to identifying the source of the capability measures, it is also important to consider how the measures were utilized in the review studies. Measure use encompasses a range of issues including: the number of manufacturing capabilities examined, the survey instrument in which the items corresponding to each measure are imbedded and the study's research context.

The measures in the review papers were typically used to assess the four standard capabilities – quality, delivery, flexibility and cost. As shown in Table 2, seven of the review papers not only examined these four capabilities but also proposed some new capabilities. For instance, the papers by Noble (1995 and 1997) divided dependability into delivery and dependability and added an additional capability called innovation.

Table 2
Studies by Type of Measure and Number of Capabilities

	Multiple-Item Measures	Developed Scaled Measures
Measured the Four Classic Capabilities	Boyer (1998) Boyer & McDermott (1999) Katuria (2000) Corbett & Whybark (2001) Boyer & Lewis (2002) Squire et al. (2006) Avella & Vazquez-Bustelo (2010) Liu et al. (2011)	Ward et al. (1995) Ward et al. (1998) Sum et al. (2004) Größler & Grübner (2006) Amoako-Gyampah & Meredith (2007) Chung & Swink (2009) Schroeder et al. (2011) Sum et al. (2012) Wu et al. (2012) Boon-itt & Wong (2016) Ehie & Muogboh (2016) Chavez et al. (2017)
Measured the Four Classic Capabilities and Added Additional Capabilities	Noble (1995) Noble (1997) Rosenzweig et al. (2003)	Swink et al. (2007) Avella et al. (2011) Wong et al. (2011) Singh et al. (2014)

Table 2 also shows that eleven of the review papers used survey instruments that relied on multiple items to measure each capability while sixteen featured survey instruments based on comprehensive capability scales. Development of such scales requires identifying a set of questionnaire items for each capability that not only adequately measures the capability but also ensures that it is distinct from the other capabilities. This condition is referred to as measurement cohesiveness. Twelve of the review papers addressed the issue of measurement cohesiveness (Table 3).

Table 3
Cohesiveness and Reliability Assessment

Author (Year)	Initial Cohesiveness	Goodness of Fit	Reliability	Number of Final Scale Items
Noble (1995)				
Ward et al. (1995)	§		§	§
Noble (1997)				
Boyer (1998)			§	§
Ward et al. (1998)	§		§	§
Boyer & McDermott (1999)				
Kathuria (2000)	§		§	§
Corbett & Whybark (2001)				
Boyer & Lewis (2002)				
Rosenzweig et al. (2003)			§	§
Sum et al. (2004)			§	§
Größler & Grübner (2006)	§		§	§
Squire et al. (2006)			§	§
Amoako-Gyampah & Meredith (2007)	§		§	§
Swink et al. (2007)			§	§
Chung & Swink (2009)	§		§	§
Avella & Vasquez-Bustelo (2010)	§		§	§
Avella et al. (2011)	§		§	§
Liu et al. (2011)	§		§	§
Schroeder et al. (2011)	§		§	§
Wong, et al. (2011)	§		§	§
Sum et al. (2012)		§	§	§
Wu et al. (2012)			§	
Singh et al. (2014)			§	§
Boon-itt & Wong (2016)			§	
Ehie & Muogboh (2016)	§		§	
Chavez et al. (2017)			§	§

The items used in the survey instrument must also exhibit internal consistency which is defined as the strength of the construct. Reliability analysis is used to determine the internal consistency of the capability measures. Cronbach's alpha is the most widely used measure for reliability analysis (Cronbach, 1951; Peter, 1979; Hinkin, 1995). Twenty-two of the studies reported the assessment of reliability using Cronbach's alpha (Table 3). All but one of these studies reported the use of alphas above the generally accepted minimum level of 0.60 suggested in the literature (Nunnally, 1978; Jones and James, 1979). The number of items in the final capability measures varied from a low of two to a high of seven items.

Once the issue of reliability has been addressed, the validity of the survey instrument should be considered. Validity is a measure of the "extent to which the instrument captures what it is intended to capture" (Hensley, 1999, p. 352). Examining validity involves assessing its components: content validity, construct validity, criterion-related validity, convergent validity, discriminant validity and predictive validity (Churchill, 1979; Hensley, 1999).

Content validity determines whether the items included in the measure capture enough of the complexity of the construct to provide a usable measure (Carmines and Zeller, 1979). Researchers support claims of content validity using some combination of the following: (1) conducting a thorough literature review, (2) pre-testing the instrument in an actual company, (3) having people outside the field examine the results of the initial factor analysis to define the factors and (4) conducting an item analysis of the measurement scales (Ahire, et al., 1996).

Although all the papers provided a literature review, only six of the studies specifically mentioned content validity (Table 4). Face validity is related to content validity but it only "incorporates variables of importance to management and portrays the direction and magnitude of their relations in a manner consistent with management logic" (Slinkman and Hanna, 1985, p. 16). Table 4 shows that two studies reported on face validity instead of content validity.

Table 4
Validity Assessment

	Validity						
Author (Year)	**Face**	**Content**	**Construct**	**Criterion-Related**	**Convergent**	**Discriminant**	**Predictive**
Noble (1995)							
Ward et al. (1995)			§	§		§	
Noble (1997)							
Boyer (1998)							
Ward et al. (1998)				§	§	§	
Boyer & McDermott (1999)							
Kathuria (2000)							
Corbett & Whybark (2001)							
Boyer & Lewis (2002)							
Rosenzweig et al. (2003)							
Sum et al. (2004)							
Größler & Grübner (2006)					§	§	
Squire et al. (2006)							
Amoako-Gyampah & Meredith (2007)			§		§	§	
Swink et al. (2007) [1]				§			§
Chung and Swink (2009)					§	§	
Avella & Vasquez-Bustelo (2010)							
Avella et al. (2011) [1]		§	§	§	§	§	
Liu et al. (2011)							
Schroeder et al. (2011)		§			§	§	
Wong et al. (2011) [1] [2]		§		§	§	§	
Sum et al. (2012) [2]	§				§	§	
Wu et al. (2012)		§	§		§	§	
Singh et al. (2014) [1]	§	§			§	§	
Boon-itt & Wong (2016) [1]					§	§	
Ehie & Muogboh (2016)					§	§	
Chavez et al. (2017) [1]		§	§			§	

Construct validity is the extent to which the measure actually represents the theoretical construct it was intended to capture (Carmines and Zeller, 1979; Churchill, 1987). There is no direct measure for construct validity but Spector (1992) suggests that researchers combine the following methods to assess it: (1) adoption of previously used measurement scale items; (2) careful item checking; and (3) exploratory factor analysis to identify unidimensional of the measurement scales (Kim and Mueller, 1978; Spector, 1992). Only five studies assessed construct validity; these papers cited comprehensive literature reviews (Ward et al., 1995), previously used scales (Amoako-Gyampah and Meredith, 2007) and factor analysis (Avella et al., 2011; Wu et al., 2012) to support claims of construct validity (Table 4).

Criterion-related validity examines the relationship between the construct measure and some "surrogate measure" of the construct (Hensley, 1999, p. 354). If a significant relationship exists between the construct and the surrogate measure, then the scale possesses criterion-related validity (Spector, 1992; Hensley, 1999). As Table 4 shows, five studies assessed criterion-related validity. These papers used varying approaches: (1) development and application of a cross-validation index based on the "use of various environmental measures to predict operations strategy" (Ward et al., 1995); (2) analysis of variance showing significant relationships between competitive priorities and choice of process (Ward et al., 1998); (3) correlations between competitive priorities and less perceptual measures (Swink et al. 2007); (4) correlations between capability measures and performance measures (Avella et al., 2011) and (5) correlations between capabilities and operational performance outcomes (Wong et al., 2011).

Convergent validity is assumed when different items in the construct measure are highly related to each other (Churchill, 1987; Spector, 1992). Measurement usually involves correlation matrices (Churchill, 1987; Spector, 1992) or confirmatory factor analysis (Ahire et al., 1996). Table 4 shows that 12 studies assessed convergent validity.

Discriminant validity assesses whether the construct measures are separate from each other (Campbell and Fiske, 1959; Spector, 1992). Discriminant validity is usually measured by conducting a χ^2 difference test (Hensley, 1999). Table 4 shows that discriminant validity was explored in 14 studies.

Predictive validity is used to compare the construct measure to a criterion measure "administered at a later point in time" (Mislevy and Rupp, 2012, p. 1077). This does not necessarily require a longitudinal study merely that the "criterion is obtained sometime after the test is given" (Lakshmi and Mohideen, 2013, p. 2756). Only one study (Swink et al., 2011) claimed predictive validity citing the correlation of study variables with the manufacturing capability variables as the basis for the claim (Table 4).

The framing of the Likert scale for the questionnaire items provides insight into how the items are used in the survey instrument. The choice of the Likert scale frame requires care because different studies may use the same items but use a different Likert scale frame. As Table 5 shows, the Likert scale measure frames varied considerably. Nine of the studies used a Likert scale frame that asked respondents to rate their use of the capability items relative to competitors and three surveys asked respondents to rate the emphasis placed on the capability questionnaire item. Two of the studies did not provide a Likert scale frame. All but two of the studies used an odd number of points, either five or seven.

Table 5
Likert Scale Design

Author (Year)	Likert Scale Frame	Points
Noble (1995)	Various, differed within capabilities	
Ward et al. (1995)	Degree of emphasis	5
Noble (1997)	Various, differed within capabilities	
Boyer (1998)	How important is the ability to	4
Ward et al. (1998)	Importance in selling the products in your primary product line	5
Boyer & McDermott (1999)	How important is the ability to	4
Kathuria (2000)	Degree of emphasis	5
Corbett & Whybark (2001)	Two types of measures: Comparison to competition and direct measures (only have points for comparison to competition)	5
Boyer & Lewis (2002)	How important is the ability to	7
Rosenzweig et al. (2003)	Organization's capability related to competitors	5
Sum et al. (2004)	Current strength/performance compared to their major competitors	7
Größler & Grübner (2006)	Amount of change over the last three years	5
Squire et al. (2006)	Performance compared to major competitors	5
Amoako-Gyampah & Meredith (2007)	Degree of emphasis that your manufacturing plant places	7
Swink et al. (2007)	Performance relative to principal competition	7
Chung & Swink (2009)	Performance relative to principal competition	7
Avella & Vasquez-Bustelo (2010)	Strategic relevance or weight	5
Avella et al. (2011)	Strategic relevance or weight	5
Liu et al. (2011)	Capability level	5
Schroeder et al. (2011)	How your plant compares with competition in your industry, on a global basis	5
Wong et al. (2011)		5
Sum et al. (2012)	Performance relative to competitors	7
Wu et al. (2012)	Extent to which used	7
Singh et al. (2014)	Performance compared with competitor	7
Boon-itt & Wong (2016)	Meet customer needs	5
Ehie & Muogboh (2016)		
Chavez et al. (2017)	Business' actual manufacturing capabilities	7

The arrangement of the questionnaire items and reverse scoring can affect survey results (Flynn et al., 1990; Alreck and Settle, 1985). Three of the studies stated that the items were grouped together on the survey in a proposed order (Ward et al., 1995; Ward et al., 1998; Schroeder et al., 2011). Only one study (Amoako-Gyampah and Meredith, 2007) reported the use of reverse scoring of some of the questionnaire items.

The research context of the study also plays a role in the use of capability measures. Context includes the purpose of the study and the setting in which the survey instrument was administered.

The review papers may be divided into two major groups defined by their purpose (Table 6). The first group used capability measures to test and compare various theoretical models of manufacturing capability. The second group of papers did not propose a new model but instead applied capability measures in studies of organizational performance.

Table 6
Study Purpose

1. Model Testing	**Study**	**Findings**
	Noble (1995)	Support for a modified sandcone: quality/dependability/delivery/cost/ flexibility/innovation
	Boyer and Lewis (2002)	Tradeoffs occur: delivery has a negative Correlation with flexibility and quality. No Support for sandcone.
	Größler and Grübner (2006)	Only a partial progression: quality/delivery. Cost and flexibility are mutually exclusive.
	Amoako-Gyampah and Meredith (2007)	Support for a modified sandcone: quality/cost/delivery/flexibility.
	Avella, et al. (2011)	Support for proposed order and added Environmental protection as fifth capability
	Schroeder, et al., (2011)	Did not support sandcone: mediated model: half of firms do not use sandcone, not just one Progression.
	Sum et al. (2012)	Tested cumulative model and found that quality is first but companies develop delivery, flexibility and cost simultaneously.
	Wu et al. (2012)	Firms can use multiple paths in building capabilities to differentiate themselves from their competitors.
	Singh et al. (2012)	No tradeoffs used. Support for threshold model (one capability is high and others are average) or average for all capabilities. Some firms had low levels for all capabilities.
	Boon-itt and Wong (2016)	No tradeoffs used. Cumulative models tested and quality/delivery/cost/flexibility is best fit.
2. Capability Applications	**Study**	**Findings**
	Ward, et al. (1995)	High performers use different strategies (capabilities) than low performers
	Noble (1997)	High performance firms tend to utilize (exploit) multiple capabilities. Sand cone progression varies by setting.
	Boyer (1998)	Infrastructural investment has a significant relationship with cost, quality and delivery.
	Ward, et al., (1998)	Significant relationship between the use of capabilities and process choice
	Boyer and McDermott (1999)	Inconsistencies in manufacturing priorities
	Kathuria (2000)	Different manufacturing types emphasize different capabilities
	Corbett and Whybark (2001)	Sandcone supported for the four groups
	Rosenzweig, et al. (2003)	High supply chain integration produces high quality, delivery reliability, process flexibility and lower cost. "Competitive capability use leads to superior business performance."
	Sum, et al., (2004)	Variety of performance measures. High performers on a variety of measures use various performance measures
	Squire, et al. (2006)	Customization requires a trade-off between delivery and cost.
	Swink et al. (2007)	Product-process integration is associated with improved quality, delivery and flexibility.

	Chung and Swink (2009)	High use firms tend to use more capabilities. Group with highest usage have tradeoffs between cost and quality; other groups tend to use more capabilities as AMT usage increases.
	Avella and Vazquez-Bustelo (2010)	Use of the four capabilities and fifth, environmental protection leads to higher sales turnover and ROI.
	Liu, et al. (2011)	Companies using strategic time orientation, supply chain integration and advanced manufacturing technology practices have higher levels of combinative competitive capabilities.
	Wong et al. (2011)	Significant relationship between capability performance and supply chain integration; high environmental uncertainty impacts capability performance.
	Ehie and Muogboh (2016)	Environmental factors (business cost, labor availability, competitive hostility and environmental dynamism) and government policies and adopted manufacturing practices have significant impact on manufacturing capabilities.
	Chavez et al. (2017)	Manufacturing capabilities related to performance is moderated by entrepreneurial orientation.

The review studies were conducted in a variety of settings ranging from single country to multiple country settings (Table 7). Only three studies limited their data to single manufacturing sectors: discrete parts (Ward et al., 1998), consumer products (Rosenzweig et al., 2003) and automotive (Wong et al., 2011). Fourteen of the studies chose the plant as the level of analysis while only three studies focused on the corporate level. Five studies targeted the company level. There was only one longitudinal study (Boyer, 1998). Table 7 also shows that 16 studies identified the use of single respondents; four identified the use of multiple respondents and seven studies did not identify whether single or multiple respondents were surveyed.

Table 7
Study Setting

Investigators (Year Published)	Countries/Industries	Level of Analysis	Participants	Respondents
Noble (1995)	Various	Plant	Various	Single
Ward et al. (1995)	Singapore/Various	Corporate	Top Exec.	Single
Noble (1997)	Various	Plant	Various	Single
Boyer (1998)	U.S./Various	Plant	Upper level mfg.	Longitudinal
Ward et al. (1998)	U.S./Discrete parts	Plant	Top Mgmt.	Multiple
Boyer & McDermott (1999)	U.S./Various	Plant	Upper level mfg.	Multiple
Kathuria (2000)	U.S./Various	Company	Mfg. mgr./GM	Multiple
Corbett & Whybark (2001)	Various		Sr. mfg. exec./GM	
Boyer & Lewis (2002)	U.S./Various	Plant	Upper level mfg. & AMT oper.	Multiple
Rosenzweig et al., (2003)	Various/Consumer	MBU	Sr. mfg. exec.	Single
Sum et al. (2004)	Singapore/Various	Corporate	CEO	Single
Größler & Grübner (2006)	Various	Plant	Oper. Dir.	Single
Squire et al. (2006)	UK/Various	Plant	Prod. Head	Single
Amoako-Gyampah & Meredith (2007)	Ghana/Various	Plant	OM Mgr.	Single
Swink et al. (2007)	North America/Various	Plant	Plant Mgr.	Single
Chung & Swink (2009)[1]	U.S./Various	Plant	Plant Mgr.	Single
Avella & Vasquez-Bustelo (2010)	Spain/Various	Firm	Plant Mgr.	Single
Avella et al. (2011)	Spain/Various	Firm	Plant Mgr.	Single
Liu et al. (2011)[1]	Various			
Schroeder et al. (2011)	Various	Plant	Plant Mgr.	Single
Wong et al. (2011)	Thailand/Automotive	Corporate	Various	Single
Sum et al. (2012)	Asia-Pacific/Various	Company	Top Mgr.	
Wu et al. (2012)	US/Various	Plant	Oper. Mgr.	
Singh et al. (2014)	Various	Plant		
Boon-itt & Wong (2016)	Thailand/Various	Company	Top Mgr.	Single
Ehie & Muogboh (2016)	Nigeria/Various	Company		
Chavez et al. (2017)	China/Various	Company	Top Managers	Single

ITEM COMMONALITY

The questionnaire items used to measure quality, delivery, flexibility and cost in the review papers were analyzed to identify commonalities. Item commonality can be thought of as the extent to which questionnaire items corresponding to a capability are used by the review studies to measure the capability.

As Table 8 shows, the most frequently used questionnaire item for quality was performance (used by 70.4% of the studies) followed by conformance and reliability (both used by 66.7% of the studies). Durability was used in 33.3% of the studies while 25.9% of the studies used features as a questionnaire item. As Table 9 shows, 18 of the review papers included additional items to measure quality. The papers by Noble (1995; 1997) and Amoako-Gyampah and Meredith (2007) each used six additional items. In total, there were 47 additional items.

Table 8
Commonality of Measurement Items

Measure	Item	% of Studies			
		Above 70%	**50 – 69%**	**40 – 49%**	**25 – 39%**
Quality	Performance	.704			
	Conformance		.667		
	Reliability		.667		
	Durability				.333
	Features				.259
Delivery	Delivery Speed	.704			
	Lead Time			.444	
	Delivery Reliability			.407	
Flexibility	Volume Changes	.777			
	Mix Changes		.630		
	Product Mix/New Product			.444	
	Range			.407	
	Customization				.333
	Fast New Product Introduction				
Cost	Overhead Costs			.444	
	Labor Productivity				.333
	Unit Costs				.259

Table 9
Questionnaire Items for Quality

Author (Year)	Performance	Conformance	Reliability	Durability	Features	Additional Items
Noble (1995)						6
Ward et al. (1995)		§	§		§	1
Noble (1997)						6
Boyer (1998)	§	§	§			0
Ward et al. (1998)	§	§	§	§		2
Boyer & McDermott (1999)	§	§	§			0
Kathuria (2000)		§				2
Corbett & Whybark (2001)						6
Boyer & Lewis (2002)	§	§	§			0
Rosenzweig et al. (2003)	§	§	§	§		1
Sum et al. (2004)	§	§				1
Größler & Grübner (2006)		§	§			0
Squire et al. (2006)	§	§		§		2
Amoako-Gyampah & Meredith (2007)	§				§	6
Swink et al. (2007)	§	§	§	§	§	0
Chung & Swink (2009)	§	§	§	§		1
Avella & Vazquez-Bustelo (2010)	§	§	§	§	§	0
Avella, et al. (2011)	§	§	§	§	§	0
Liu et al. (2011)	§	§	§	§		0
Schroeder et al. (2011)	§	§				0
Wong et al. (2011	§		§			2
Sum et al. (2012)	§	§	§			1
Wu et al. (2012)		§	§	§	§	1
Singh et al. (2014)	§				§	1
Boon-itt & Wong (2016)	§		§			2
Ehie & Muogboh (2016)			§			4
Chavez et al. (2017)	§		§			2
Number of Items	**19**	**18**	**18**	**9**	**7**	**47**

Only three items used to measure delivery displayed at least moderate commonality. Delivery speed was used in 70.4% of the studies (Table 8). Lead time was used in 44.4% of the studies while delivery reliability was used in 40.7%. Twenty-four papers used additional items to measure delivery. In total there were 50 additional items (Table 10).

Table 10
Questionnaire Items for Delivery

Author (Year)	Delivery Speed	Lead Time	Delivery Reliability	Additional Items
Noble (1995)	§		§	0
Ward et al. (1995)	§		§	2
Noble (1997)	§		§	0
Boyer (1998)	§	§		2
Ward et al. (1998)		§		4
Boyer & McDermott (2007)	§	§		2
Kathuria (2000)	§	§		1
Corbett & Whybark (2001)				5
Boyer & Lewis (2002)	§	§		1
Rosenzweig et al. (2003)			§	1
Sum et al. (2004)		§		2
Größler & Grübner (2006)	§	§	§	0
Squire et al. (2006)	§	§	§	1
Amoako-Gyampah & Meredith (2007)	§			1
Swink et al. (2007)	§			3
Chung & Swink (2009)	§			3
Avella & Vazquez-Bustelo (2010)	§	§		1
Avella, et al. (2011)	§	§		1
Liu et al. (2011)			§	1
Schroeder et al. (2011)	§			1
Wong et al. (2011		§	§	3
Sum et al. (2012)	§			1
Wu et al. (2012)				3
Singh et al. (2014)	§			3
Boon-itt & Wong (2016)	§		§	3
Ehie & Muogboh (2016)	§		§	2
Chavez et al. (2017)		§	§	3
Number of Items	**19**	**12**	**11**	**50**

There was more commonality among items used to measure flexibility than among the items for any of the other capabilities (Table 8). Five items were identified as commonly used: volume changes (77.7%), mix changes (63%), product mix/new product range (44.4%), customization (40.7%) and fast new product introduction (33.3%). Fifteen papers included additional flexibility items; there were 30 such items (Table 11).

Table 11
Questionnaire Items for Flexibility

Author (Year)	Volume Changes	Mix Changes	Product Range/Mix	Customization of Products	Fast/New Product Introduction	Additional Items
Noble (1995)	§		§	§		2
Ward et al. (1995)					§	3
Noble (1997)	§		§	§		2
Boyer (1998)	§	§				1
Ward et al. (1998)	§			§	§	1
Boyer & McDermott (1999)	§	§				1
Kathuria (2000)		§		§	§	2
Corbett & Whybark (2001)						5
Boyer & Lewis (2002)	§	§				1
Rosenzweig et al. (2003)	§	§	§			0
Sum et al. (2004)	§	§	§		§	0
Größler & Grübner (2006)	§	§				0
Squire et al. (2006)						4
Amoako-Gyampah & Meredith (2007)		§				1
Swink et al. (2007)	§		§	§		1
Chung & Swink (2009)	§		§	§		1
Avella & Vazquez-Bustelo (2010)	§	§	§		§	0
Avella, et al. (2011)	§	§	§		§	2
Liu et al. (2011)	§	§		§		0
Schroeder et al. (2011)	§	§			§	0
Wong et al. (2011	§	§	§	§		0
Sum et al. (2012)	§	§		§	§	0
Wu et al. (2012)	§		§			0
Singh et al. (2014)	§	§				0
Boon-itt & Wong (2016)	§	§	§	§		0
Ehie & Muogboh (2016)					§	3
Chavez et al. (2017)	§	§	§	§		0
Number of Items	**21**	**17**	**12**	**11**	**9**	**30**

The most commonly used items to measure cost were: overhead costs (44.4%), labor productivity (33.3%) and unit costs (25.9%) (Table 8). Twenty-six papers proposed additional items for cost. Noble's (1995; 1997) and Corbett and Whybark's (2001) studies each used seven additional items. As Table 12 shows, 65 additional items were used for the cost capability.

Table 12
Questionnaire Items for Cost

Author (Year)	Overhead Costs	Labor Productivity	Unit Costs	Additional Items
Noble (1995)	§			7
Ward et al. (1995)	§		§	2
Noble (1997)	§			7
Boyer (1998)		§		1
Ward et al. (1998)		§		2
Boyer & McDermott (1999)		§		1
Kathuria (2000)		§		2
Corbett & Whybark (2001)				7
Boyer & Lewis (2002)		§		1
Rosenzweig et al. (2003)			§	1
Sum et al. (2004)			§	1
Größler & Grübner (2006)	§	§		2
Squire et al. (2006)				5
Amoako-Gyampah & Meredith (2007)	§			2
Swink et al. (2007)	§			1
Chung & Swink (2009)	§			1
Avella & Vazquez-Bustelo (2010)		§		2
Avella, et al. (2011)		§		3
Liu et al. (2011)				1
Schroeder et al. (2011)			§	0
Wong et al. (2011	§			3
Sum et al. (2012)		§	§	1
Wu et al. (2012)	§		§	1
Singh et al. (2014)				3
Boon-itt & Wong (2016)	§			3
Ehie & Muogboh (2016)	§		§	2
Chavez et al. (2017)	§			3
Number of Items	**12**	**9**	**7**	**65**

Table 13 summarizes the number of commonly used items for each of the four capabilities and also reveals the pervasiveness of dissimilar items. The ratio of common items to dissimilar items for each capability was small, ranging from a high of 16.7% for the flexibility items to a low of 4.7% for the cost items.

Table 13
Questionnaire Item Analysis

Capability	Number of Items Commonly Used	Number of Items Not Commonly Used	Ratio of Common Items to Dissimilar Items
Flexibility	5	30	16.67%
Quality	5	47	10.64%
Delivery	3	50	6.00%
Cost	3	65	4.7%

ADDITIONAL ANALYSIS

Review results discussed in Section 3 reveal considerable variation in the extent to which the studies performed validity checks. Review findings also reveal differences in the sources of the capability measures, number of capabilities measured, the use of comprehensive scaling and

the purpose of the study. These results suggest an additional line of inquiry: to what extent, if any, do measure sourcing, number of capabilities, the use/non-use of comprehensive scales and the study purpose affect the use of validity checks? This general line of inquiry can be stated more formally as a series of research questions:

Question 1: Do the studies that developed their own capability measures differ in their use of validity checks from those that used measures from other studies or databases?

Question 2: Do the studies that measured the four standard capabilities differ in their use of validity checks from those studies that measured the standard four as well as additional capabilities?

Question 3: Do studies that used comprehensive scales to measure capabilities differ in their use of validity checks from those studies that relied solely on multiple-item measures?

Question 4: Do studies that tested capability models differ in their use of validity checks from those studies that did not test a capability model but rather applied capability measures to examine performance?

One-tailed t-tests were used to examine these four questions. Table 14 shows that papers that developed their own measures had significantly more validity checks than papers that adopted measures from previous studies (p-value = 0.0030). Similarly, studies that used comprehensive scales had a significantly greater number of validity checks than papers relying on multi-item measures of capability (p-value < 0.0001). Likewise, papers that tested capability models had more validity checks than application papers (p-value = 0.0111). In contrast, the number of capabilities measured did not make a significant difference in the number of validity checks undertaken (p-value = 0.1984).

Table 14
Extent of Validity Checks
T-Test Results[1]

Question	Average	p-value
1. Source of Measure		
Developed for Study	2.31	0.0030
Previously Developed	1.07	
2. Number of Capabilities		
Original Four	1.50	0.1948
Original Four Plus Additional	2.14	
3. Type of Measure		
Comprehensive Scaled	2.37	<0.0001
Multiple-Item	2.25	
4. Purpose		
Test Model	2.60	0.0111
Application	1.12	

[1] All data was checked for equal variances prior to conducting t-tests. Unequal variance tests were used as needed.

The review results in Section 4 reveal extensive variation in item commonality. The results also confirm that a few high commonality items did occur in the review studies. This paper will define a high commonality item as a questionnaire item that was used by at least 60% of the review papers. There were only six high commonality items: performance, conformance and reliability, which were used to measure quality; delivery speed, which was used to measure delivery and volume changes and mix changes, which were used to measure flexibility (see Table 8). Given the variability in measure sourcing and measure use in the review papers, yet another line of inquiry emerges: To what extent, if any, do measure sourcing, number of capabilities, the use/non-use of comprehensive scales and the study purpose affect the use of

high commonality items? This particular line of inquiry can be stated more formally as a series of research questions:

Question 5: Do the studies that developed their own capability measures differ in their use of high commonality items from those that used measures from previous studies or databases?

Question 6: Do the studies that measured the four standard capabilities differ in their use of high commonality items from those studies that measured the standard four as well as additional capabilities?

Question 7: Do studies that used comprehensive scales to measure capabilities differ in their use of high commonality items from those studies that relied solely on multiple-item measures?

Question 8: Do studies that tested capability models differ in their use of high commonality items from those studies that did not test a capability model but rather applied capability measures to examine performance?

Proportion tests were used to examine this series of questions. The following procedure was followed for each of the four questions. First, the review papers were divided into two groups based on the particular aspect of measure sourcing or measure use under consideration. Second, the number of instances of high commonality item use were counted each of the two groups. Third, total counts of all items used for each of the two groups were tabulated. Fourth, the proportion for high commonality items for each of the two groups was computed (Table 15). Finally, a proportion test was run to determine whether the proportions of the two groups were significantly different.

Table 15
Proportion Tests for Item Commonalities

Question	Proportion	p-value
5. Source of Measure		
Developed for Study	0.3027	0.3062
Previously Developed	0.2569	
6. Number of Capabilities		
Original Four	0.2927	0.2979
Original Four Plus Additional	0.2414	
7. Type of Measure		
Comprehensive Scaled	0.1940	0.0053
Multiple-Item	0.3280	
8. Purpose		
Test Model	0.3237	0.1362
Application	0.2538	

Table 15 shows that the proportion of high commonality items in papers that used comprehensive scales was significantly greater the proportion of high commonality items in papers that relied on multi-item measures of capability (p-value = 0.0053). In contrast, neither the source of the measures (p-value = 0.3062) nor the number of capabilities measured (p-value = 0.2979) made a significant difference in the use of high commonality items. Likewise, the purpose of the study made no difference in the use of high commonality items (p-value = 0.1362).

DISCUSSION OF REVIEW RESULTS

Two key findings result from this review: 1) only six of the items used by the studies could be described as high commonality items and 2) there is a considerable lack of consensus in measure sourcing, Likert scale framing and use of validity checks.

The paucity of high commonality items in the review papers is striking because it demonstrates a genuine lack of consensus about which items best measure each of the four traditional capabilities. This implies that different researchers often stress different aspects of each capability making it difficult to establish one consistent measure for each capability. This situation can impede progress in both theory development and practice. It is worth noting that the use of comprehensive measurement scales was associated with a greater extent of high commonality items than the use of simpler multiple-item measures. This outcome is not surprising since comprehensive scale development requires rigorous assessment of scale cohesiveness, reliability and validity.

In contrast to comprehensive scaling, the variation that occurred in measure sourcing did not significantly affect high commonality item use. Approximately half the papers took measures from previous research – including large scale databases. Such databases do allow the researcher to examine many different types of companies over large geographic areas and may thus make it easier to generalize study findings; however, they do not guarantee high item commonality. Papers that developed their own measures had as much high item commonality as papers that used measures from previous research.

Variation in Likert scale framing for questionnaire items was considerable. Only 33% of the frames were based on comparison of the company's capabilities with those of competitors. Other frames occurred with much lower frequencies. While many papers utilized previously developed measures, the item frames they selected often differed from the original frames. Changing item frames can be problematic since an item frame provides the respondent with a context for answer choices and thus influences how the respondent perceives the questionnaire items. Altering the respondent's perceptions can affect the results of the survey questionnaire.

The review papers also exhibited substantial variation in the number of validity checks undertaken. T-tests revealed that papers that papers using comprehensive scales used significantly more validity checks than papers using multi-item measures. One would anticipate such a finding since a comprehensive scale development process entails rigorous assessment of initial scale construction, including evaluation of scale validity (Schwab, 1980; Hensley, 1999). Researchers typically use a combination of validity measures to conduct such an evaluation. A second t-test revealed that papers that developed their own measures used significantly more validity checks than papers using measures from previous studies. All but two papers that used original measures also utilized comprehensive scales; thus, this finding mirrors the result of the first t-test. A third t-test showed that model testing studies had significantly more validity checks than application papers. Half the model testing papers used comprehensive scales which may account for the relatively high number of validity checks in this group.

Although each of the papers that utilized comprehensive scales did perform a combination of validity checks, no paper addressed all types of validity. Only one paper (Swink et al., 2007) examined predictive validity. This finding confirms Boyer and Pagell's (2000) assertion that predictive validity remains a relatively ignored topic in manufacturing capability.
Another relatively ignored topic is longitudinal measurement of manufacturing capability – despite calls for longitudinal studies in the literature (Sarmiento et al., 2010). Only Boyer (1998) adopted a longitudinal approach to capability measurement while examining the relationship between the capabilities and advanced manufacturing technologies. Boyer (1998) surveyed plants in the United States, gathering data in 1994 and again in 1996.

Taken together, the review results discussed above suggest a number of ideas for improving capability measurement. These recommendations are delineated in the following section.

RECOMMENDATIONS

Several recommendations on capability measurement stem from the review results. These recommendations encompass a range of issues including measure sourcing, item selection, Likert scale framing, reliability, validity checks and measurement timing.

In addressing the issue of measure sourcing, the researcher should begin by considering the purpose of the study. If the purpose is to develop and test a theoretical capability model, the researcher should, as a preliminary step, examine the capability measures that were used to in conjunction with such theoretical models in the past. The researcher may wish to consult the first section of Table 6 for a list of theoretical capability models. In contrast, if the purpose of the study is the application of capability measures in a specific industry or manufacturing firm, then the researcher may wish to consult section two of Table 6 which summarizes papers that applied capability measures in studies of organizational performance. The researcher may also wish to refer to Table 7 to determine the countries or industries that comprised the application settings. Those papers with a research setting that matches the one the researcher plans to use merit close attention.

Once a preliminary examination of relevant papers is complete, the researcher must decide whether to adopt capability measures used by a previous study or to develop original measures. The results of this review showed that either approach to measure sourcing can be used for theoretical modeling papers or for application papers. Both approaches to measure sourcing share some key methodological considerations; however, the process of measure development entails some additional issues.

When developing original capability measures, the researcher must choose the items that will comprise each of the individual capability measures. There should be at least three items per measure but the total number of items per capability should be kept as small as possible. This approach will limit the length of the questionnaire, which is desirable since response rates tend to increase as the survey length decreases (Cronbach and Meehl, 1955).

Since each measure will consist of a small number of items, the researcher must choose them carefully. Measures used in previous studies are an obvious source of individual items for new measures. The researcher can consult Table 8 for a listing of the most commonly used items for each measure. In addition, the researcher can consult Tables 9 through 12 which show the extent to which each review paper in this study used the most common items for measuring quality, delivery, flexibility and cost. Selecting items with high commonality contributes to more uniform ways of measuring each capability, which supports both theoretical studies as well as application papers. Alternatively, the researcher may decide to devise new items or select rarely used items from previous studies rather than using commonly used items for each measure. In this case, the researcher must be able to provide a clear rationale why these particular items were chosen.

Regardless of the degree of commonality of the items selected, the researcher must be able to demonstrate a clear link between the items used to measure a specific capability and the capability construct. Hinkin (1995) recommends "a strong theoretical framework and employing a rigorous sorting process that matches items to construct definitions." The researcher may wish to enlist academics or practitioners with expertise in capability

management to carry out the sorting process. This type of sorting process will help to establish the content validity of the items that are chosen.

Once the items are selected, a Likert scale must be chosen to frame the items. The researcher should proceed with caution when using measures from multiple papers for item sourcing. The frames from previous measures may differ from paper to paper. In choosing a single frame for the questionnaire items, the researcher must be sure that replacing the original frame will not radically alter how a respondent would perceive the item. Since it is difficult to achieve consensus in capability measurement when item interpretation varies study by study, the researcher should consider choosing a frequently used frame for the Likert frame. Comparison of a company's capabilities with those of competitors constitutes one such possibility as it was used by more of the review papers than any other framing choice. It also comports with Sarmiento et al.'s (2010) recommendation of an external reference for measurement. This review showed that there are other possibilities for framing choice including degree of emphasis, strategic relevance or weight and importance specific firm abilities. No matter which frame is finally selected, it should match the purpose of the study.

After the items and their framing have been determined, the researcher should consider a pre-test of the questionnaire. Participants can be chosen from the group of intended survey respondents or from individuals outside the target population. This initial survey testing helps the researcher gauge the instrument before it is administered to the ultimate respondents.

The preceding discussion illustrates that there are a number of methodological considerations to address when creating new capability measures instead of relying on previously used measures. However, both approaches to measure sourcing have other methodological concerns in common.

The first concern is the reliability of the proposed measures. While this study showed that almost all the review papers reported Cronbach's alpha to assess reliability, it is important to note that "Cronbach's alpha is a minimum measure of reliability" (Hensley, 1999, p.352). The researcher may wish to follow the example of Sum et al. (2012) and also conduct an item analysis for each measure. Item analysis provides a way of checking the accuracy of the measure's Cronbach's alpha (Spector, 1992; Hensley, 1999). It is an iterative process that requires the calculation of both item-remainder coefficients and Cronbach's alpha coefficients at each step (Spector, 1992). The process concludes when a suitable Cronbach's alpha level is achieved. According to Spector (1992, p. 32) the value 0.70 is a "widely accepted rule of thumb" although Nunnally (1978) suggests that 0.60 is acceptable for "exploratory research" (Black and Porter, 1996).

There are other ways to support Cronbach's alpha in addition to item analysis. These include the split-halves test recommended by Black and Porter (1996), the comparison of at least two factor analysis rotations and the Werts-Linn-Jorsekog coefficient (Ahire et al., 1996; Black and Porter, 1996; Hensley, 1999).

A second methodological concern involves validity. Validity checks should be thorough. Researchers should try to include checks for several types of validity, even those that cannot be measured directly. For instance, content validity is subjective but can be established through a combination of extensive literature reviews, expert sorting of proposed items for each measure, pre-tests of the instrument in actual manufacturing contexts and, in the case of scaled measures, outside reviews of initial factor analysis results (Hensley, 1999). There is also no

direct measure for construct validity; it too can be established through a comprehensive literature review and the use previously used scales and factor analysis.

A final methodological consideration involves longitudinal capability measurement which was utilized by only one of the review papers. A longitudinal measurement approach does demand more time and effort than gathering data at a single point in time but can yield benefits. A longitudinal approach allows the researcher to investigate changes in how the capabilities are perceived over time and can be helpful in identifying what Schroeder et al. (2011, p. 4886) called the "long lasting effects" of adhering to a particular capability model. In addition, the researcher could expand a longitudinal measurement process to include the tracking of operational data. Operational data could thus supplement the perceptual data gained from scaled questionnaires and provide a fuller picture of manufacturing capability at the firm(s) in the study.

This review has demonstrated that much work that remains to be done in the measurement of manufacturing capabilities. Continued improvement of capability measurement is crucial because good measures remain essential for both practitioners and researchers. Practitioners need good measures because capability measurement constitutes a key component of any audit of a firm's manufacturing strategy and helps in making strategic decisions. Good capability measurement also can also advance theory and help researchers build on existing models. Simply put, model development begins with good measures for capability constructs and good measures begin with a sound measurement development process.

References

Ahire, S.L., D.Y Golhar, Waller, M.A., 1996. Development and validation of TQM implementation constructs. *Decision Sciences* 27, 1, 23-56.

Alreck, P., Settle, R., 1985. *The Survey Research Handbook*. Richard D. Irwin Company, Homewood IL:

Amoako-Gyampah, K., Meredith, J.R., 2007. Examining cumulative capabilities in a developing economy. *International Journal of Operations and Production Management* 27, 9, 928- 950.

Anderson, J.C., Cleveland, G., Schroeder, R.G., 1989. Operations strategy: literature review. *Journal of Operations Management* 8, 2, 135-158.

Avella, L., Vazquez-Bustelo, D., 2010. The multidimensional nature of production competence and additional evidence of its impact on business performance. *International Journal of Operations and Production Management* 30, 6, 548-583.

Avella, L, D., Vazquez-Bustelo, D., Fernandez, E., 2011. Cumulative manufacturing capabilities: an extended model and new empirical evidence. *International Journal of Production Economics* 3, 1, 707-729.

Black, S.A., Porter, L.J., 1996. Identification of the critical factors of TQM. *Decision Sciences* 27 (1), 1-21.

Boer, H., Holweg, M., Kilduff, M., Pagell, M., Schmenner, R., 2015. Making a meaningful contribution to theory. *International Journal of Operations and Production Management* 35, 9, 1231-1252.

Boon-itt, S., Wong, C.Y., 2016. Empirical investigation of alternate cumulative capability models: a multi-method approach. *Production Planning & Control* 27, 4, 299-311.

Boyer, K.K., 1998. Longitudinal linkages between intended and realized operations strategies. *International Journal of Operations and Production Management* 18, 4, 356-373.

Boyer, K.K., McDermott, C., 1999. Strategic consensus in operations strategy. *Journal of Operations Management* 17, 3, 289-305.

Boyer, K.K., Pagell, M., 2000. Measurement issues in empirical research: improving measures of operations strategy and advanced manufacturing technology. *Journal of Operations Management* 18, 3, 361-374.

Boyer, K.K., Lewis, M.W., 2002. Competitive priorities: investigating the need for trade-offs in operations strategy. *Production and Operations Management* 11, 1, 9-20.

Bronzo, M., Valadares de Oliveira, M.P., McCormack, K., 2012. Planning, capabilities, and performance: an integrated value approach. *Management Decision* 50, 6, 1001-1021.

Campbell, D.T., Fiske, D.W., 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56, 81-105.

Carmines, E.G., Zeller, R.A., 1979. *Reliability and Validity*. Sage Publications, Beverly Hills, CA.

Chavez, R., Yu, W., Jacobs, M.A., Feng, M., 2017. Manufacturing capability and organizational performance: the role of entrepreneurial orientation. *International Journal of Production Economics* 184, 33-46.

Churchill, G.A., 1979. A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research* 16, 2, 64-73.

Churchill, G.A., 1987. *Marketing Research: Methodological Foundations, Fourth Edition*. The Dryden Press, Chicago, IL.

Chung, W., Swink, M., 2009. Patterns of advanced manufacturing technology utilization and manufacturing capabilities. *Production and Operations Management* 18, 5, 533-545.

Corbett, C., Whybark, D.C., 2001. Searching for the sandcone in the GMRG data. *International Journal of Operations and Production Management* 21, 7, 965- 980.

Cronbach, L.J., 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 297-334.

Cronbach, L.J., Meehl, P.E., 1955. Construct validity in psychological tests. *Psychological Bulletin* 52, 4, 281-302.

Ehie, I., Muogboh, O., 2016. Analysis of manufacturing strategy in developing countries. *Journal of Manufacturing Technology Management* 27, 2, 234-260.

Fisher, M., 2007. Strengthening the empirical base of operations management. *Manufacturing & Service Operations Management* 9, 4, 368-382.

Flynn, B.B., Sakakibara, S., Schoeder, R.G., Bates, K.A., Flynn, E.J., 1990. Empirical research methods in operations management. *Journal of Operations Management* 9, 2, 250-284.

Größler, A., Grübner, A., 2006. An empirical model of the relationships between manufacturing capabilities. *International Journal of Operations and Production Management* 26, 5, 458-485.

Hayes, R.H., Wheelwright, S.C., 1984. *Restoring Our Competitive Edge*, Wiley, New York.

Hensley, R.L., 1999. A review of operations management studies using scale development techniques. *Journal of Operations Management* 17, 3, 343-358.

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

Ibrahim, S.E., 2010. An alternative methodology for formulating an operations strategy: the case of BTC-Egypt. *Management Decision* 48, 6, 868-893.

Jones, A.P., James, L.R., 1979. Psychological climate: dimensions and relationships of individual and aggregated work environment perceptions. *Organizational Behavior and Human Performance* 23, 201-250.

Kathuria, R., 2000. Competitive priorities and managerial performance: a taxonomy of small manufacturers. *Journal of Operations Management* 18, 6, 627-641.

Kim, J.Q., Mueller, C.W., 1978. *Introduction to Factor Analysis*. Sage Publications, Newbury Park, CA.

Lakshumi, S., Mohideen, M.A., 2013. Issues in reliability and validity of research." *International Journal of Management Research and Review* 3, 4, 2752-2758.

Liu, N., Roth, A.V., Rabinovich, E., 2011. Antecedents and consequences of combinative competitive capabilities in manufacturing. *International Journal of Operations and Production Management* 31, 12, 1250-1286.

Meredith, J.R., 1998. Building operations management theory through case and field research. *Journal of Operations Management* 16, 4, 441-454.

Meredith, J.R., 2009. Issues in the modeling-empiricism gap. *Journal of Supply Chain Management* 45, 1, 44-48.

Minor, E.D., Hensley, R.L., Wood, D.R., 1994. A review of empirical manufacturing strategy studies. *International Journal of Operations and Production Management* 14, 1, 5-25.

Mislevy, J.L., Rupp, A.A., 2012. Predictive validity. In *Encyclopedia of Research Design*, edited by Neil J. Salkind, 1077-1078. Sage Publications, Inc., Newbury Park, CA.

Noble, M.A., 1995. Manufacturing strategy: testing the cumulative model in a multiple country context. *Decision Sciences* 26, 5, 693-721.

Noble, M.A., 1997. Manufacturing competitive priorities and productivity: an empirical study. *International Journal of Operations and Production Management* 17, 1, 85-99.

Nunnally, J.C., 1978. *Psychometric Theory*. McGraw-Hill, New York.

O'Regan, N., Ghobadian, A., 2004. The importance of capabilities for strategic direction and performance. *Management Decision* 42, 2, 292-312.

Peter, J.P., 1979. Reliability: a review of psychometric basics and recent marketing practices. *Journal of Marketing Research* 16, 2, 6-17.

Rosenzweig, E.D., Easton, G.S., 2010. Tradeoffs in manufacturing? A meta-analysis and critique of the literature. *Production and Operations Management* 19, 2, 127-141.

Rosenzweig, E.D., Roth, A.V., Dean, J.W., 2003. The influence of an integration strategy on competitive capabilities and business performance: an exploratory study of consumer products manufacturers. *Journal of Operations Management* 21, 4, 437-456.

Roth, A.V., Miller, J.G., 1992. Success factors in manufacturing. *Business Horizons* 35, 4, 73-81.

Sarmiento, R., Sarkis, J., Byrne, M., 2010. Manufacturing capabilities and performance: a critical analysis and review. *International Journal of Production Research* 48, 5, 1267-1286.

Schmenner, R.W., Swink, M.L., 1998. On theory in operations management. *Journal of Operations Management* 17, 1, 97-113.

Schroeder, R.G., Shah, R., Peng, D.X., 2011. The cumulative capability 'sand cone' model revisited: a new perspective for manufacturing strategy. *International Journal of Production Research* 49, 16, 4875- 4901.

Schwab, D.P., 1980. Construct validity in organizational behavior. In *Research in Organizational Behavior, Vol. 2*, edited by Staw, B.M., Cumings, L.L. JaI Press, Greenwich, CT.

Singh, P.J., Wiengarten, F., Nand, A.A., Betts, T., 2014. Beyond the trade-off and cumulative capabilities models: alternative models of operations strategy. *International Journal of Production Research* 53, 13, 4001-4020.

Skinner, W., 1969. Manufacturing – missing link in corporate strategy. *Harvard Business Review* May-June, 136-145.

Slinkman, C. W., Hanna, M.E., 1985. Ingredients of a good regression based model. *The Journal of Business Forecasting Methods & Systems* 4, 3, 16-18.

Spector, P.E., 1992. *Summated Rating Scale Construction: An Introduction*. Sage University Paper Series on Quantitative Application in the Social Sciences, Series No. 07-082, Sage, Newbury Park CA.

Squire, B., Brown, S., Readman, J., Bessant, J. 2006. The impact of mass customization on manufacturing trade-offs. *Production and Operations Management* 15, 1, 10-21.

Sum, C., Kow, L.S., Chen, C., 2004. A taxonomy of operations strategies of high performing small and medium enterprises in Singapore. *International Journal of Operations and Production Management* 24, 3, 321-345.

Sum, C.C., Singh, P.J., Heng, H.Y., 2012. An examination of the cumulative capabilities model in selected Asia-Pacific countries. *Production Planning & Control* 23, 10-11, 735-753.

Swink, M., Narasiman, R., Wang, C., 2007. Managing beyond the factory walls: effects of four types of strategic integration on manufacturing plant performance. *Journal of Operations Management* 25, 148-164.
Vickery, S.K., Dröge, C.L.M., Markland, R.E., 1993. Production competence and business strategy: do they affect business performance? *Decision Sciences* 24, 2, 435-455.

Wacker, J.G., 1998. A definition of theory: research guidelines for different theory-building research methods in operations management. *Journal of Operations Management* 16, 4, 361-385.

Ward, P.T., Duray, R., Leong, G.K., Sum, C.C., 1995. Business environment, operations strategy, and performance: an empirical study of Singapore manufacturers. *Journal of Operations Management* 13, 99-115.

Ward, P.T., McCreery, J.K., Ritzman, L.P., Sharma, D., 1998. Competitive priorities in operations management. *Decision Sciences* 29, 4, 1035-1046.

Wong, C.Y., Boon-itt, S., Wong, C.W.Y., 2011. The contingency effects of environmental uncertainty on the relationship between supply chain integration and operational performance. *Journal of Operations Management* 29, 6, 604-615.

Wu, S.J., Melnyk, S.A., Swink, M., 2012. An empirical investigation of the combinatorial nature of operational practices and operational capabilities. *International Journal of Operations and Production Management* 32, 2, 121-155.