Technical Synopsis of the Comp-XM® Assessment

January 2015

Author: Erich Dierdorff

EXECUTIVE SUMMARY

This report describes the technical specifications of Capsim's Comp-XM®. In brief, Comp-XM is a simulation-based assessment tool designed to broadly capture "business acumen" – knowledge and skills related to the various functions of organizational management. Comp-XM goes beyond measuring recall of acquired knowledge assessed by standardized exams in a static testing environment. Comp-XM instead provides both an assessment of acquired knowledge and an assessment of how effectively individuals use this knowledge to manage a business in a changing and competitive simulated market. The primary goal of this technical synopsis is to present evidence of the veracity and psychometric quality of the measures included in Comp-XM.

This report is organized into six sections. The first section provides a brief overview of Comp-XM. For more detailed information or a demonstration of Comp-XM, interested readers may directly contact Capsim at welcome@capsim.com. The second section of this report describes the procedure and results from analysis of the original set of test questions contained in Comp-XM, including information on reliability and item discrimination. The third section describes the procedure and results of a content-related validity review and the generation of parallel items for the original set of test questions. The fourth section details the development of two supplemental item packets. These additional tests were designed to measure ethical decision-making and teamwork competence. The fifth section describes the procedure and results from analysis of the revised test question set as well as the two supplemental packets. Finally, the sixth section discusses future development and research avenues for Comp-XM.

TABLE OF CONTENTS

OVERVIEW OF THE COMP-XM ASSESSMENT

Comp-XM is a simulation-based assessment tool designed to broadly capture what could be generally termed as "business acumen." In short, business acumen encompasses more specific knowledge and skills related to the various functions of business or organizational management (e.g., finance/accounting, operations/production, marketing, human resources, and so forth). These common functions of business are well recognized in schools of business and are not only reflected in the common structure of business schools (i.e., departments within the institution), but also in various standards published by accrediting bodies such as the Association to Advance Collegiate Schools of Business (AACSB), Accreditation Council for Business Schools and Programs (ACBSP), and International Assembly for Collegiate Business Education (IACBE).

One of the challenges of assessing business acumen, particularly in formal training settings such as universities, is capturing individuals' capacities to apply the knowledge that they have mastered from their training or coursework. Bridging this "knowing-doing" gap in terms of assessment is impossible to do using only traditional assessment methods such as exams, essays, and papers. When properly constructed these types of conventional assessments do quite well in evaluating individuals' knowledge acquisition, but simply cannot provide trainers or educators direct information about individuals' skilled performance (i.e., successfully applying what they know).

Comp-XM goes beyond measuring recall of acquired knowledge assessed by standardized exams in a static testing environment. Comp-XM instead provides both an assessment of acquired knowledge and a picture of how effectively individuals use this knowledge to manage a business in a changing and competitive market. This testing environment assesses whether individuals can adapt to the successes and failures in the early rounds of the simulation to effectively shape their company's future. By doing so, Comp-XM evaluates functionally specific knowledge and skills, the integration of these competencies, and most importantly, an individual's overall business acumen.

Comp-XM Assessment Procedure

As a simulation-based assessment, Comp-XM implements two assessment techniques. The first technique examines the effectiveness of decisions made during the simulation itself (i.e., when individuals manage their companies). Thus, this assessment portion of Comp-XM is best characterized as measuring an individual's skills; that is, the application of knowledge or the behavioral demonstration of proficiency. The second technique uses a series of more traditional test questions that are designed to measure knowledge mastery. These items are called "board query" questions to reflect the idea that they are coming from the executive board of an individual's simulated company. Both these assessment techniques will be described in more detail below.

There are five "rounds" in the Comp-XM assessment. The first four rounds require making simulation decisions and answering sets of board query questions. The final round requires only answering board query questions. A typical Comp-XM administration takes approximately six hours to complete and can be conducted in-person or online, as well as in a single administration or across multiple testing periods.

Simulation-derived Assessment. As mentioned above, a portion of Comp-XM assessment is derived from how effectively individuals make decisions when managing their simulated company over four decision rounds. The decisions that are required fall into the common functional domains

of business/organizational management and include financial accounting, marketing, research and development, production, and human resources. Illustrative examples of the kinds of decisions made and their simulated effects in each of these areas are shown in the table below. Importantly like any real business, success in the simulation requires not only effective function-specific decisions, but also the effective integration of these decisions to align with a chosen strategy. Thus, performance also requires that individuals demonstrate strategic decision-making skills as well as cross-functional skills.

Functional Area	Decision Examples	Decision Impacts
Research and Development	Assess current products based on chosen strategy, change product specifications of those products, or introduce new products	Product attractiveness Product age Material cost Capital expenditure
Marketing	Set the product prices; determine promotional budgets; set sales budgets; set credit policies	Product attractiveness Customer awareness Customer accessibility Payment terms
Operations/production	Set production schedules; determine plant capacity requirements; set automation levels; design new product requirements	Product availability Labor costs Inventory management Future plant capacity analysis Asset & plant utilization Fixed capital investment allocation
Finance/accounting	Issue/re-purchase stock; issue short-term debt; issue/retire long-term bonds; pay shareholder dividends; balance debt portfolio; manage pro-formas	Cash on-hand Emergency loan requirements Stock price Ratio performance Interest rates
Human Resources	Fund process improvement initiatives; fund total quality management initiatives; determine training hours; set recruitment budgets; determine staffing levels	Employee productivity Employee turnover Customer demand Labor & material costs

Because the simulation requires effective cross-functional decision-making, it is important to assess the performance of the simulated company in a holistic manner. In other words, measuring the effectiveness of how an individual manages his or her company in an overall sense, as opposed to only focusing on financial performance, for example. To accomplish this aim, company performance is measured using the "Balanced Scorecard," which is well recognized in both academic and industry settings as an effective framework to capture a organization's overall performance and/or strategic progress. In short, the Balanced Scorecard assesses performance in four areas or quadrants. [1] These areas are shown below as well as the specific metrics used in Comp-XM to measure each quadrant.

[1] See Kaplan, R. S., & Norton, D. P. (1992, 1993, 1996a, 1996b, 2000). See reference list for full citations.

Balanced Scorecard Quadrant	Comp-XM Metrics
Customer Perspective How well are customers' needs being met? A focus on customer awareness, service, & satisfaction Reflected in performance measurements that show quality, time, performance, & service	Overall awareness Overall accessibility Customer survey score Accounts receivable lag Cumulative R&D cycle time Unit share in product segments
Internal Business Process Perspective How well is the business running & meeting market needs & expectations? A focus on internal business operations Reflected in performance measurements that show effectiveness & efficiency of productivity & service delivery	Contribution margin Plant utilization levels Days of working capital Stock-out costs Inventory carrying costs
Innovation & Learning Perspective How well are employees developed? A focus on employee training, attitudes, & performance Reflected in performance measurements that show individual productivity, learning, retention, & so forth	Employee turnover rate Employee productivity Material reduction R&D reduction Administrative cost reduction Customer demand increase
Financial Perspective How well is the business performing in terms of profitability & economic viability? A focus on profits, shareholder value, cash flow, & so forth	Stock price Profits Leverage (assets-to-equity)

As seen in the descriptions above, the Balanced Scorecard depicts how an individual manages his or her company in a holistic and integrative manner. This characteristic aligns with the integrative nature of the simulation, which requires individuals to make various functional decisions consistent with their chosen strategy. This broader focus of the Balanced Scorecard thus allows for a more encompassing depiction of competence in cross-functional business management.

Board Query Assessment.

This portion of the Comp-XM assessment uses a more conventional examination method, whereby individuals are required to answer multiple-choice test questions. These test items are distributed across the five rounds of Comp-XM and were written to assess aspects of functional knowledge and business strategy relevant to the simulation, particularly as they relate to each assessed learning objective being targeted.

The set of test questions includes 41 total questions. These questions span content related to strategy (5 questions), finance/accounting (18 questions), human resources (5 questions), operations (6 questions), and marketing (7 questions). Consistent with the simulation-based assessment approach of Comp-XM, these 41 questions are embedded in each individual's simulation experience. This is accomplished by using an individual's simulation decisions and associated results for their company as the basis for the exam questions that are generated. In this way, the data that populate an individual's exam are directly derived from his or her own unique simulation input. It is important to note that the same topical content is still being assessed, although the data used to generate a given item will differ across test takers. For example, a question that assesses knowledge

about calculating financial ratios would require all test takers to use the same formula to calculate return-on-equity; however, information used as input into the calculation will differ across test takers because it is based on each person's simulated environment. Thus, individuals see identical exam topics, but questions and answers are unique to the situations they create in their own simulations.

[2] See Weekly & Ployhart (2006). Situational judgment tests: Theory, management, and application. Mahwah, NJ: Erlbaum.

PHASE 1: ANALYSIS OF ORIGINAL BOARD QUERY ITEMS

Purpose

The purpose of this initial phase of Comp-XM test development was to examine the psychometric properties of the original set of default Comp-XM board query (BQ) questions. This original set contained 42 total questions. Emphasis in this phase was placed on estimating test reliability as well as levels of item difficulty and discrimination. Such estimation is important as prevailing professional standards for test developers and distributors stipulate the primary responsibility for obtaining and disseminating information about an assessment procedure's reliability.[3] Results from this analysis were used to inform test revisions.

Procedure

Sample data on the 42 default Comp-XM BQ items comprised 26,661 individuals that completed the assessment between 2011 and 2012. These individuals were students at the undergraduate and graduate levels from a variety of post-secondary educational institutions.

Reliability. The reliability of a test refers to the extent to which the test is likely to produce consistent scores. There are several types of reliability, including internal consistency, split-half, and test-retest. In general, estimates of internal consistency are most often used for item analysis. In this case, high reliability means that the questions of a test "pull together" or are internally consistent. Thus, individuals who correctly answer a given question are more likely to correctly answer other questions. Internal consistency reliability can be displayed by calculating Cronbach's coefficient alpha estimate, which varies from 0 to 1.0. Because responses were coded dichotomously (correct or incorrect), reliability was estimated using Kuder-Richardson (KR-20) estimates. Below are some are general guidelines for interpreting reliability coefficients.

Internal consistency	Interpretation
.90 – 1.0	Excellent
.80 – .90	Very good
.70 – .80	Good
.60 – .70	Somewhat low
.50 – .60	Low
< .50	Poor

* Nunnally, J. C. (1967) Psychometric Theory. New York: McGraw-Hill.

Difficulty. The number of individuals who correctly answer a particular item determines item difficulty. Item difficulty is relevant for determining whether or not individuals have learned the concept being tested. In addition, item difficulty impacts a test item's capacity to discriminate between individuals who know the material and those who do not.

For items with a single correct alternative, the item difficulty is simply the percentage of individuals who correctly answer an item. The *item difficulty index* ranges from 0 to 1.0, with higher values indicating easier questions. When there is more than one correct alternative, item difficulty is the

[3] See American Educational Research Association (1999). Standards for educational and psychological testing. Washington, DC: Author.

average score on that item divided by the highest number of points for any single response option. Two general "rules-of-thumb" about item difficulty levels:

1. Desirable difficulty levels are slightly higher than midway between chance (guessing) and perfect scores (100% correct) for a given item.
2. Items in the difficulty range of .30 to .70 tend to yield important differences between individuals' levels of knowledge.

Below are some are general guidelines for interpreting item difficulty levels. It should be noted that these "ideal" levels do not communicate high or low quality for a given item because ultimately it is the test developer that determines the preferred "mix" of item difficulty levels.

Format	"Ideal" Difficulty
Five-response multiple-choice	.70
Four-response multiple-choice	.74
Three-response multiple-choice	.77
True-false items	.85

*Lord, F.M. (1952). The Relationship of the Reliability of Multiple-Choice Test to the Distribution of Item Difficulties, Psychometrika, 18, 181-194.

Discrimination. Item discrimination refers to the capacity of a test item to differentiate among individuals on the basis of how well they know the material being tested. Item discrimination goes beyond determining the proportion of individuals who answer correctly and depicts more specifically who answers correctly. That is, item discrimination shows whether individuals who did well on the entire test also did well on a particular test item.

Two estimates were used to estimate item discrimination. The first was the *discriminability index (D)*. This index was calculated by first dividing individuals into a high scoring group and a low scoring group (top one-third and bottom one-third), then estimating the proportion of each group that answered a particular test item correctly, and finally subtracting proportion of the bottom group from the proportion for the top group.

The second index was the *discrimination coefficient (p-bis)*, which is estimated using a point-biserial correlation coefficient between performance on an item and performance on the entire test (item scores compared to total scores). This coefficient indicates the tendency of individuals selecting the correct answer to have high overall test scores. This coefficient should be positive, indicating that individuals answering correctly tend to have higher overall scores or that individuals answering incorrectly tend to have lower overall scores. It also provides an estimate of the degree to which an individual item is measuring the similar content as other items in the test. Below are some are general "rule-of-thumb" guidelines for interpreting item discrimination.

Discrimination Index (*D*)		Discrimination Coefficient (r_{bis})	
Range	Interpretation	Range	Interpretation
> .40	High	> .30	High
.20 to .40	Fair	.10 to .30	Fair
< .20	Low	< .10	Low

* R.L. Ebel (1954). Procedures for the Analysis of Classroom Tests, Educational and Psychological Measurement, 14, 352-364. Nunnally, J. C. (1967) Psychometric Theory. New York: McGraw-Hill.

Results

The correlation between performance in the simulation (Balanced Scorecard scores) and performance on the BQ items was .59, providing evidence that the two assessment techniques are capturing related, but unique information. This lends further support to the proposition that the simulation more closely measures business-related skills, while the BQ questions more proximally capture business-related knowledge.

The internal consistency reliability estimate (KR-20) for the entire test was .92, which is well within the "excellent" range. The table below provides descriptive statistics for item analyses across the other calculated indices.

	Difficulty Index	**Discrimination Index (D)**	**Discrimination Coefficient (r_{bis})**
Mean	.43	.68	.44
Standard Deviation	.08	.16	.14
Median	.43	.66	.48
Maximum	.57	.98	.65
Minimum	.23	.33	.08
f("low")	4 (10%)	0 (0%)	0 (0%)

* *f*("low") = count of items falling into the "low" category for each index, per guidelines discussed earlier.

As shown in the table above, the 42 default BQ items performed well in terms of difficulty and discrimination, with average levels all falling within "high" ranges for each index. In addition, even those items falling one standard deviation below the mean were still within the favorable range for each index.

Four items were identified with potentially unfavorable difficulty index estimates. However, examining the discrimination indices and coefficients for these items suggested that they were adequately functioning in general (i.e., *D* range = .94-.98 and r_{bis} range = .46-.52). Thus, these items were initially retained in the default set for further examination in Phase 2 discussed next.

PHASE 2: REVIEWING & AUGMENTING BOARD QUERY ITEMS

Purpose

The purpose of this second phase of Comp-XM test development was twofold: a) to examine the content-related validity of the default item set, and b) to generate parallel item sets for each of the default items.

Procedure

Conventional content review procedures were followed, which began with the identification of subject matter experts (SMEs). All SMEs were tenured full-time faculty across multiple functional areas, including finance/accounting, marketing, operations, and management. This variety in functional expertise was important because SMEs were asked to review BQ items that specifically pertained to their own functional domains. The table below lists the participating SMEs and the functional areas they reviewed.

SME	Functional area reviewed
Dr. Cindy Durtschi	Finance/accounting
Dr. Michael Tippins	Marketing
Dr. Michael Dixon	Operations/production
Dr. Erich Dierdorff	Strategy/HR

The SME review process began with "live" one-on-one demonstrations of Comp-XM. These demonstrations allowed for the SMEs to review and experience the simulation environment, as well as provided an opportunity to ask questions about the assessment. Following the demonstration, SMEs were sent a content review packet relevant to their content domain. These packets contained a general overview, detailed instructions for the rating task, and the set of BQ items to be reviewed. As an example, below are the review packet instructions and tasks sent to the finance/accounting SME.

> *The main goal of your content review of Comp-XM board query questions is to ensure adequate coverage of generalizable and foundational "business acumen" concepts that are germane to Capsim simulation performance. A secondary goal of your review is to expand the overall pool of board query questions by generating sets of parallel items that can be used interchangeably to create alternate, equivalent versions of the board query exam. Your content review entails the following four steps. More specific instructions are found in the respective sections for each step.*
>
> *Step 1: Review the support materials and instructions for your designated content area (see pages 2-3 below and attachments).*
>
> *Step 2: Review existing Comp-XM questions and rate each in terms of applicability (see pages 4-11 below).*
>
> *Step 3: Identify any obvious "gaps" in terms of concept/knowledge coverage for your designated area (see pages 12 below).*
>
> *Step 4: Write at least 3 parallel items that tap into the same concept tested by existing item (e.g., financial ratios); if applicable, write 3-5 items tapping into "gap" area items identified in step 3 (see pages 13-20 below).*

Each SME was also provided more detailed descriptions of the types of content topics found in Comp-XM that were pertinent to their functional domain. For example, for the SME review of accounting items, the description read:

Your specific content review focuses on board query questions that assess accounting knowledge within the exam. To reiterate, board query questions link back to students' performance during the simulation. Thus, below is a brief description of the types of accounting activities that students must perform during the simulation. In addition, please consult attachments A and B to gain an understanding of the types of accounting data that are made available to the students during the simulation.

From an operational perspective students' accounting decisions are primarily based on setting an accounts payable and receivable term, analyzing the cash flow, balance sheet, and income statement during each simulation round. Students are provided with a series of "Proformas" or "what-if" statements that project what their three financial statements will be if all of their sales projections are correct (see attachment C, at the end of this document). Typical accounting activities that students must perform using these data include:

- *Examining the balance sheet, cash flow and income statements of their company and their competitors*
- *Setting accounts receivable and payable terms*
- *Performing break-even analyses*
- *Examining operational scenarios that will impact a company's financial ratios*

*There are a total of seven board query questions that require the application of accounting principles. These questions are presented in the following section (Step 2 of your review). Correct answers are marked with an asterisk **

There are certain names used by the simulation. Company names in the simulation are: Andrews, Baldwin, Chester, Digby, Erie, and Ferris. Product names follow the first letter of the company name (e.g., Baldwin might have products named "Brat" and "Bam").

SMEs were asked to review each BQ item and to provide a rating of the extent to which the item represented a concept that is applicable or representative of foundational business acumen in their designated content area (e.g., marketing). SMEs were instructed that "foundational" meant the types of functional knowledge that are more generalizable rather than hyper-specific in nature; that is concepts that everyday managers might confront when running a business versus calculations that apply more specifically to particular occupations (e.g., accountants or financial analysts or marketing analysts or HR specialists). These ratings were generated using a 3-point scale (0 = not relevant, 1 = probably relevant, 2 = definitely relevant).

SME Review Results. The results of the SME ratings are shown in the table below and are grouped by content area to facilitate interpretation.

Content Area	Avg. Relevance Rating	Details	Adjusted Avg.
Accounting/finance (18 items)	2.0		
Human resources (5 items)	2.0		
Operations (6 items)	1.3	1 item was rated '0'	1.6
Marketing (7 items)	1.7		
Strategy (6 items)	1.8	1 item was rated '0'	2.0

* Adjusted average = average after removing the problematic item.

Feedback from the SME review also noted that four of the original strategy items were essentially capturing the same topic ("identifying strategies") and suggested that these be combined into a single question set (i.e., four parallel items measuring identifying strategies). This combination in conjunction with the removal of the other "irrelevant" item (noted in the table above) created room for expanding the strategy-oriented items. The strategy domain SME wrote three additional items to expand the tested strategy content. This included items measuring strategic analysis, competitive analysis, and developing mission/vision statements. Similarly, the SME for the operations domain wrote a replacement question for the original item judged as "irrelevant" (noted in the table above). This new item tested content related to effective and efficient manufacturing plant utilization.

The SME review process resulted in a new default BQ item set of 41 total questions. The topics covered in each functional domain in this default set are shown in the table below.

Content Area	Topics Covered
Finance/accounting (18 items)	Break Even Analysis Calculating Book Value Calculating Dividend Yield Calculating Ratios from the Annual Report Calculating Simple Ratios Calculating Stock Repurchase Cash Management DuPont Analysis Effect of Investment Decisions Effects of Change in Depreciation Expense on the Financial Statements Identifying Change in Equity Identifying Fixed vs. Variable Costs Interpreting the Cash Flow Statement Retirement of Debt Revenue Recognition Understanding the Accounting Equation Understanding the Accounting Equation Understanding the Carrying Values of Items on the Balance Sheet
Human resources (5 items)	Calculating Future Labor Wages Calculating Productivity Impact Calculating Recruiting Costs Calculating Separation Costs Calculating Training Costs
Operations (6 items)	Operational Impact of Unit Margin Total Quality Management Break Even Analysis Capacity Analysis Determining Acceptable Inventory Levels Cost of Right-Sizing Plant Plant Utilization
Marketing (7 items)	Creating Marketing Budgets Demand Analysis Forecasting Identifying Competitors Using the "Four P's" Identifying Marketing Efficiency Identifying Price Elasticity Market Sizing
Strategy (5 items)	Competitive Analysis Developing Mission/Vision Statements Identifying Strategies Identifying Tactics for Building Competitive Advantage Strategic Analysis

SME Parallel Item Generation. The SMEs were also asked to write parallel or equivalent items for the default BQ questions within their respective domains of expertise. SMEs were instructed that these new questions should be interchangeable with the "parent" question in order to provide a pool of items that assess a given concept (i.e., the same underlying knowledge/skill principle). On average nearly three parallel items were written for each question in the default BQ list (Mean = 2.73). The table below summarizes the number of parallel items that were generated for each of the content areas tested in the default question list.

Content Area	Num. of "Parent" Items	Num. of Parallel Items
Accounting/finance	18	66
Human resources	5	14
Operations	6	9
Marketing	7	12
Strategy	5	11
Total	**41**	**112**

PHASE 3: DEVELOPMENT OF SUPPLEMENTAL ITEM SETS

Purpose

The purpose of this third phase of Comp-XM test development was to create supplemental item sets that extend beyond the default BQ question set. The need for these additional item sets was driven by academic customers' desire to have "built-in" questions assessing other learning goals typical to business school accreditation requirements. Two specific content areas were the focus of this phase: a) teamwork effectiveness and b) ethical decision-making.

Procedure

An evidence-based approach was taken to develop two situational judgment tests (SJTs) capturing competencies related to teamwork effectiveness and ethical decision-making. Drs. Erich Dierdorff and Robert Rubin directed the SJT development process. Briefly, SJTs are low fidelity simulations that present a respondent with a written description of a work situation and multiple responses to the situation. The respondent is instructed to endorse one choice only. SJTs are different than traditional multiple-choice tests. Unlike standardized multiple-choice tests, SJTs do not emphasize assessment of "declarative knowledge" (i.e. knowing about something). Instead, SJTs measure "procedural knowledge," which reflects knowing how to do something. Research evidence has supported a strong relationship between SJTs and performance in a given role, primarily because procedural knowledge is a more direct precursor to actual behavior than declarative knowledge.[4]

Two general approaches can be employed when developing SJTs. One is to take a critical incident approach that emphasizes the identification of particular situations and then designates alternative actions in those situations. Another tactic is a construct-oriented approach that emphasizes the particular dimensions that underlie performance in a given domain and then identifies situations and actions reflecting those dimensions. The latter, construct-oriented approach was employed to design the current SJTs for teamwork effectiveness and ethical decision-making because of the large body of scholarship on these content areas. Conventional procedures were employed to design and develop the SJT.[5] Below are the general steps followed in the development of the two SJTs.

1. Reviewed theoretical and empirical literatures to develop conceptual frameworks to be assessed by the SJT.
2. Generated multiple clusters of situations that aligned with this framework.
3. Reviewed situations for redundancy, consistency, complexity, and representativeness of each framework category.
4. Developed response options to each situation.
5. Created initial pool of SJT items and reviewed the items for clarity and consistency.
6. Presented initial items to subject matter experts, who reviewed the items for clarity and representativeness.
7. Pilot tested initial items with a student sample (graduate and undergraduate students in a school of business).
8. Analyzed SME responses and feedback, as well as pilot test results
9. Revised the initial SJT items to reflect SME feedback and/or pilot test results.
10. Created a final version of the SJT, with a core set of items and an additional pool of equivalent items.

---

[4] See Chan, D. & Schmitt, N. (2002). Situational judgment and job performance. Human Performance, 15, 233-54. Lievens, F., Peeters, H., Schollaert, E. (2008). Situational tests: A review of recent research. Personnel Review, 37, 426-441. McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. Personnel Psychology, 60, 63–91.

[5] See Weekley, J. A., Ployhart, R. E. (2006). Situational judgment tests: Theory, management, and application. Mahwah, NJ: Erlbaum

Teamwork Competence SJT. Following the steps described, a framework consisting of nine dimensions of teamwork was derived. These dimensions span both individual- and team-level constructs. Below is a summary of these dimensions and associated references that support their importance to effective teamwork.

Level	Dimensions & Definitions	Sources [6]
Individual	**Performance Preparation** – analyzing and planning activities that enable performance	Bowers et al (1993); Gladstein (1984); Marks et al. (2001); Mathieu et al. (2000); Prince & Salas (1993); Rousseau et al. (2006); Salas et al. (2005); Stevens & Campion (1994); Weldon & Weingart (1993); Weldon et al. (1991)
	Performance Execution – collaborating and accomplishing team tasks	Bowers et al (1993); Campion et al. (1993); Cannon-Bowers et al. (1995); Cohen (1994); Dominick et al. (1997); Erez et al. (2002); Kozlowski & Bell (2003); Marks et al. (2001); Mathieu et al. (2000); McIntyre & Salas (1995); Morgan et al. (1986); Smith-Jentch et al. (1998); Rousseau et al. (2006); Stevens & Campion (1994); Tannenbaum et al. (1992); Tesluk & Mathieu (1999)
	Performance Monitoring – assessing and evaluating ongoing performance	Cannon-Bowers et al. (1995); Marks et al. (2001); McIntyre & Dickson (1992); McIntyre & Salas (1995); Rousseau et al. (2006); Salas et al (2005); Stevens & Campion (1994); Weldon et al. (1991)
	Performance Adjustment – adapting and generating performance improvements	Bowers et al (1993); Cannon-Bowers et al. (1995); Janz et al. (1997); Marks et al. (2001); Price & Salas (1993); Salas et al. (2005); Stevens & Campion (1994); Rousseau et al. (2006)
Team	**Team Confidence** – belief that team has the capability to perform successfully	Bandura (1997); DeRue et al. (2010); Dierdorff & Ellington (in press); Gist (1987); Gully et al. (2002); Kirkman et al. (2004); Lindsley et al. (1995); Marks et al. (2001)
	Team Cooperation – high-quality, reciprocal exchange of information among team members	Campion et al. (1993); Dierdorff et al. (2011); Eby & Dobbins,(1997); Erez et al. (2002); Kozlowski & Bell (2003); Mathieu et al. (2000); Seers (1989; Seers et al. (1995); Taggar & Brown (2001)
	Team Coordination – integration of individual efforts toward team goals	Cannon-Bowers et al. (1995); Hinsz et al. (1997); Marks et al. (2001); Rousseau et al. (2006); Salas, Stagl, et al. (2004); Salas et al. (2005); Stewart & Manz (1995); Zaccaro et al. (2001)
	Team Cohesion – shared commitment and affinity toward the team	Bandow (2001); Beal et al. (2003); Kirkman et al. (2004); Salas et al. (2005); Spreitzer et al. (1999); Webber (2002)
	Team Conflict –disagreement and friction among team members	DeDreu & Weingart (2003); Jehn (1997); Marks et al. (2001); Rousseau et al. (2006); Stevens & Campion (1994); Taggar & Brown (2001); Tannenbaum et al. (1992)

[6] See reference section for full citations

Drs. Dierdorff and Rubin created an initial pool of 33 items. These items were reviewed for clarity and consistency, and presented to SMEs. Three SMEs were asked to participate in the development of the SJT (see table below). Each of these individuals has specific expertise in the team performance/team training domain, including published research and organizational consulting.

Teamwork SJT SMEs
Dr. Suzanne Bell
Dr. Kemp Ellington
Dr. David Fisher

The SMEs were first asked to complete the 33-item SJT themselves. Below is a brief summary of results from the SME responses.

- Majority of the items (64%) had unanimous correct responses across the 3 SMEs
- Most of the remaining items (30%) had only a single incorrect response across the 3 SMEs
- Only 2 items had two (of three) incorrect responses across the 3 SMEs

After completing the SJT, SMEs were asked to independently review the 33 items, which were organized by the nine dimensions with associated definitions also provided. More specifically, SMEs were asked to rate each item with respect to its adequacy in representing its respective teamwork dimension. As common in test development, the SJT items were rated on a 3-point scale of construct representativeness. (1 = not representative, 2 = somewhat representative, and 3 = representative). SMEs were also asked to provide feedback for improvements. A "content validity ratio" (CVR) was computed from the SME ratings of representativeness to aid interpretation of the review.[7] CVR values range from +1 to -1, with positive values indicating that at least half the SMEs rated the item as representative. Below is a brief summary of the results from the SME ratings.

- Majority of SJT items (51%) received unanimous rating of "3"
- Most of the remaining items (42%) received ratings of two "3s" and a single "2"
- Only 2 items were identified as potentially problematic (negative CVR values)

The 33-item SJT was next administered to students from two management courses (1 graduate-level and 1 undergraduate-level). The usable sample size ranged from 30 to 34 students depending upon the specific SJT item. Below is a brief summary of the findings from the pilot test.

- "Ceiling effects" were not evident in that the highest score produced was 85% correct.
- Results indicated the SJT was "moderate" in level of overall difficulty:
    - Average scores were 69% for graduate students and 62% for undergraduate students

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

- A majority of the initial SJT items performed well in terms of difficulty and discrimination:
  - Average percentage of correct responses across the 33 items was 65%, with a standard deviation of 20%
  - Least difficult item = 93% correct responses
  - Most difficult item = 18% correct responses

Results from the SME panel review and the pilot test were used to inform final revisions to the 33 items. Based on specific SME feedback, two items were revised to improve their representativeness of the underling teamwork effectiveness dimension. Other revisions to the instrument primarily concerned any items where only very small percentage of students in the pilot study chose a given response option. Here, revisions focused on improving the response options by making the incorrect response options less obvious (i.e., making the incorrect distracter options more "distracting").

The final teamwork effectiveness SJT question set consisted of 9 "core" items. Each of the nine teamwork effectiveness dimensions is equally represented in these core items. A reserve pool of an additional 24 items was also created for purposes of generating alternate equivalent forms of the SJT. Efforts were made to ensure similar levels of difficulty across the core items and the reserve pool items. For example, the average percentage of correct responses from the pilot study across the 9 core items was 65.3%, compared to an average percentage of 64.2% across the 24 reserve pool items.

Ethical Decision-making SJT. Following the steps earlier in this report section, a framework describing the common stages of an ethical decision-making process was developed. Critical to any assessment in the domain of ethics is an understanding of the operational definition of ethics, ethical issues and ethical decision-making. An *ethical issue* is one where a person's actions, when freely performed, may harm or benefit others.[8] Therefore, *ethical decision making* is the reasoning process involved in determining courses of action and necessarily involves consequences for others and volition on the part of the decision maker. A decision is deemed *ethical* when it is both legal and morally acceptable to the larger community and is based upon careful consideration of the facts.[9]

In developing an SJT for business ethics and ethical decision-making, two distinct literatures were consulted. First, philosophical traditions have examined the nature of moral principles through moral discourse. Ethical philosophers such as Emanuel Kant, John Rawls and John Stuart Mill are some examples of well-known philosophers who have embedded theories of utilitarianism, universalism, and distributive justice into our daily ethical decision making discourse. Second, a growing literature on business ethics exists which is both descriptive and empirical, and provides ample frameworks and data to offer a prescriptive approach to making ethical decisions. These literatures were used to derive key dimensions along a prescriptive ethical decision-making framework.

Most critical is that the SJT assesses procedural knowledge regarding the appropriate steps to take or decision-points to consider in an ethical decision making framework. It does not assess the ethicality

[8] See Jones, T. M. (1991). Ethical decision making by individuals in organizations: An issue contingent model. Academy of Management Review, 15, 366-395.
[9] See Hartman, L. P., & DesJardins, J. R. (2008). Business ethics: Decision-making for personal integrity and social responsibility. McGraw-Hill Irwin.

of a given decision nor does it assess respondents' moral reasoning ability based on Kohlberg's stages of moral development, for example. Put simply, the SJT assesses an individual's capability to engage in the process of ethical decision-making.

Upon review of the relevant literature on ethical decision-making, a framework describing four stages of an ethical decision-making process was developed. The table below shows these four stages as well as definitions for each. [10]

Category/Stage of Ethical Decision-making	Definition
I. Investigate Ethical Issues	Perceive the hallmarks of an ethical dilemma or problem. Gather relevant facts related to, but not limited to: legal issues, precedent in the organization, authority to make a decision, organization rules, policies or regulations governing the decision. Use ethical theory/perspectives to frame or define the ethical issue. Expose problem to common philosophical traditions. Use categories of common ethical problems to situate the issue such as "bribery", "abuse of resources", "conflict of interest", "discrimination," etc.
II. Identify Primary Stakeholders	Identify individuals, groups and/or organizations that are likely to be affected by any decision. Differentiate between primary and secondary stakeholders. Determine central obligations to each group of primary stakeholders.
III. Increase Alternative Courses of Action	Expand solution set to three or more alternatives. Use effective decision-making techniques such as brainstorming or affinity grouping to promote the quantity of decision.
IV. Inspect Consequences of Alternatives	Examine alternative against ethical criteria including equity, equality and needs-based procedural justice perceptions as well as feasibility, risks, and relationships. Determine the cost/benefits to each alternative and respects or violations of universal rights. Use prevailing standards of ethicality and "checks" to prevent personal bias or gain.

Drs. Dierdorff and Rubin created an initial pool of 25 SJT items. These items were reviewed for clarity and consistency, and presented to SMEs. Two SMEs were asked to participate in the development of the SJT (see table below), both have experience researching and teaching ethical decision-making in the broader domain of organizational behavior.

Ethical Decision-making SJT SMEs
Dr. Laura Parks
Dr. Alyssa Westring

SMEs were asked to independently review the initial 25 SJT items, which were organized by the four stages of decision-making with associated dimension definitions also provided. More specifically,

[10] Stages derived from sources such as Janis & Mann (1977); Rest (1985); Trevino (1986); Ferrell & Gresham (1985); Hunt & Vitell (1986); Jones (1991); McDevitt, Giapponi, & Tromley (2006); Wilhelm (2006); Johnson (2007); Hartman & Des Jardins (2008); Trevino & Nelson (2010); Baldwin, Bommer, & Rubin (2012). See reference section for full citations.

SMEs were asked to rate each item with respect to its adequacy in representing its respective ethical decision making category/stage. The SJT items were rated on a 3-point scale of representativeness. (1 = not representative, 2 = somewhat representative, and 3 = representative). SMEs were also asked to provide feedback for improvements. Below is a brief summary of the results from the SME ratings.

- 50% (12 items) received unanimous rating of "3" across the two SMEs
- Most of the remaining items (40%) received ratings of "3s" and "2s" across the two SMEs
- Only 3 items were identified as potentially problematic.

The 25-item SJT was administered to students from two management courses (1 graduate-level and 1 undergraduate-level). The usable sample was 40 students. Below is a brief summary of the findings from the pilot test.

- "Ceiling effects" were not evident in that the highest score produced was 80% correct.
- Results indicated the SJT was "moderate to high" in level of overall difficulty:
  - Average score = 51% correct, with a standard deviation of 10%
  - Highest score = 80% correct
  - Lowest score = 32% correct
- A majority of the initial set of SJT items performed well in terms of difficulty and discriminability:
  - Least difficult item = 93% correct responses
  - Most difficult item = 23% correct responses

Results from the SME panel review and the pilot test were used to inform final revisions to the 25 items. Based on specific SME feedback, one item was deleted leaving a total pool of 24 items. A total of 9 items were revised to improve their representativeness of the assigned decision-making stage and/or to improve the response options by making the incorrect response options more or less "distracting."

The final SJT instrument consisted of 10 core items. Because some of the categories are larger in nature the categories are not balanced with equal numbers of questions. Specifically, stages I and IV have four associated questions, whereas stages II and III have two items each. A reserve pool of an additional 14 items was created for purposes of generating alternate equivalent forms of the SJT. Efforts were made to ensure similar levels of difficulty across the core items and the reserve pool items. For example, roughly 50% of the pilot sample received greater than 50% of the correct responses from both the core items and reserve items.

PHASE 4: ANALYSIS OF NEW DEFAULT & SUPPLEMENTAL BOARD QUERY ITEMS

Purpose

The purpose of this phase of Comp-XM test development was to examine the psychometric properties of the new default set of 41 BQ questions, the teamwork and ethical decision-making SJT question sets, as well as any parallel items with sufficient data. Similar to Phase 1, emphasis was placed on estimating test reliability as well as levels of item difficulty and discrimination. Results from this analysis were used to inform test revisions.

Procedure

Sample data on the revised set of 41 default Comp-XM BQ items comprised 26,166 individuals that completed the assessment between January 2013 and April 2014. These individuals were students at the undergraduate and graduate levels from a variety of post-secondary educational institutions. Operational sample sizes differed depending on the item. For the 41 "parent" questions the operational samples sizes ranged from 15,752 to 26,166 (M = 17,883). For the parallel items to these parent questions, the operational sample sizes varied from a low of 243 to a high of 5596 (M = 1869). Although some of the parallel items had relatively smaller operational samples, the vast majority (93%) had at least 500 respondents.

The same general procedure followed in Phase 1 was again applied in this phase. Estimates of reliability, difficulty, and discrimination were again examined.

Results

The correlation between performance in the simulation (Balanced Scorecard scores) and performance on the BQ questions was .51. The internal consistency reliability estimate (KR-20) for the entire test (default questions + two supplemental SJTs = 60 questions) was .93. The reliability estimate for only the default items that measure functional content areas of business was .91. Both of these estimates are well within the acceptable range. The tables below provide descriptive statistics for item analyses across the other calculated indices for both the entire battery of 180 items (all parent items and currently implemented parallel items) and for only the 152 default test items.

Total Item Battery (k=180)	**Difficulty Index**	**Discrimination Index (D)**	**Discrimination Coefficient (r_{bis})**
Mean	.60	.56	.41
Standard Deviation	.17	.16	.11
Median	.61	.59	.43
Maximum	.95	.89	.64
Minimum	.10	.07	.01
f("low")	9 (5%)	2 (1%)	3 (2%)

* f("low") = count of items falling into the "low" category for each index, per guidelines discussed earlier.

Default Item Battery (k=152)	Difficulty Index	Discrimination Index (D)	Discrimination Coefficient (r_{bis})
Mean	.61	.56	.42
Standard Deviation	.13	.14	.10
Median	.61	.59	.44
Maximum	.95	.82	.64
Minimum	.19	.10	.07
f("low")	6 (4%)	1 (1%)	1 (1%)

* f("low") = count of items falling into the "low" category for each index, per guidelines discussed earlier.

As shown in the table above, the full BQ item bank performed well in terms of difficulty and discrimination, with average levels all falling within "high" ranges for each index. The default BQ item bank performed similarly well in terms of difficulty and discrimination. Finally, even those BQ items falling one standard deviation below the mean were still within the favorable range for each index in both sets of analyses.

These results also suggested that 3 of the items from the 180-item pool warranted possible revision. This conclusion was based on the items' performance across the difficulty and discrimination estimates. The items are described below.

Item Content	Item Type	Results	Revision Decision
F-261: Effects of Change in Depreciation Expense on the Financial Statements	Parallel	r_{bis} = .07 D = .10 Difficulty = .18	Delete item; three other parallels already exist for this topic
I-524: Promoting Team Confidence	Parent	r_{bis} = .15 D = .01 Difficulty = .09	Retain item; one incorrect response option was too "distracting"; option was revised to reduce distraction
E-591: Identifying Primary Stakeholders	Parent	r_{bis} = .07 D = .15 Difficulty = .17	Retain item; one incorrect response option was too "distracting"; option was revised to reduce distraction

FUTURE DEVELOPMENT

Test development is a process of continual evidence-based improvement. Toward this end, there are numerous potential development projects that could further enhance assessment quality, expand assessment scope, as well as examine relationships of Comp-XM performance with other instruments and criteria.

With regard to further enhancing assessment quality, additional psychometric analysis is central to this effort. This would include reanalyzing response data for several of the parallel items as well as the supplemental question packets where the current samples sizes were relatively smaller compared to the default set of Comp-XM questions. In addition, further analysis is needed to examine the performance of the item revisions described in the previous section. Such examinations can ensure that these revised items are performing as expected.

The current Comp-XM assessment could also be expanded in at least two ways. First, developing additional parallel items would allow for a greater number of equivalent forms of the assessment. Second, developing additional supplemental question packets would offer the opportunity to expand the assessment content of the current topical scope of Comp-XM. For instance, many users are academic clients concerned with assurance of learning for accreditation purposes. The extent to which Comp-XM covers the key learning goals that generalize across these institutions that greater the utility of the assessment. For non-academic clients, additional question packets that align with key managerial competencies or more specific functional knowledge could provide similar utility.

Finally, examining relationships between performance in Comp-XM and in other assessments could provide further insight into the knowledge and skills being measured as well as validity-related evidence. For example, investigating relationships between commonly used knowledge tests, such as the ETS Major Field Tests for business, could provide more evidence of convergent-related validity. Examining relationships between Comp-XM and other performance outcomes, such as overall grade point averages in academic environments and job performance in work settings, could provide valuable evidence pertaining to criterion-related validity.

REFERENCES (not fully included in footnotes)

Baker, D. F. (2008). Peer assessment in small groups: A comparison of methods. Journal of Management Education, 33. 183-209.

Baldwin, T. T., Bommer, W. H., & Rubin, R. S. (2012). Managing organizational behavior: What great managers know and do. Boston, MA: McGraw-Hill/Irwin.

Bandow, D. (2001). Time to create sound teamwork. The Journal for Quality and Participation, 24, 41-47.

Bandura A. (1997). Self-efficacy: The exercise of control. New York: W.H. Freeman.

Beal, D. J., Cohen, R. R., Burke, M. J., & McLendon, C. L. (2003). Cohesion and performance in groups: A meta-analytic clarification of construct relationships. Journal of Applied Psychology, 88, 989-1004.

Bowers, C. A., Morgan, B. B., Salas, E., & Prince, C. (1993). Assessment of coordination

Brooks, C. M. & Ammons, J. L. (2003). Free riding in group projects and the effects of timing, frequency, and specificity of criteria in peer assessments. Journal of Education for Business, 78, 268-272.

Brutus, S. & Donia, M. B. L. (2010). Improving the effectiveness of students in groups with a centralized peer evaluation system. Academy of Management Learning & Education, 9, 652-662.

Campion, M. A., Medsker, G. J., & Higgs, A. C. (1993). Relations between work group characteristics and effectiveness: Implications for designing effective work groups. Personnel Psychology, 46, 823-850.

Cannon-Bowers, J. A., Tannenbaum, S. I., Salas, E., & Volpe, C. E. (1995). Defining competencies and establishing team training requirements. In R. A. Guzzo & E. Salas (Eds.), Team effectiveness and decision making in organizations (pp. 333-381). San Francisco: Jossey-Bass.

Cascio, W. F. & Aguinis, H. (2005). Applied psychology in human resources management (6th Ed). Upper Saddle River, NJ: Prentice Hall.

Cohen, S. G. (1994). Designing effective self-managing work teams. In M. M. Beyerlein & D. A. Johnson (Eds.), Advances in interdisciplinary studies of work teams: Theories of self-managed work teams (pp. 67-102). London: JAI.

Davies, W. M. (2009). Groupwork as a form of assessment: Common problems and recommended solutions. Higher Education, 58, 563-584.

Davis, D. et al. (2010). Assessing team member citizenship in a capstone engineering design course. International Journal of Engineering Education, 26, 1-13.

DeDreu, C. K. W., & Weingart, L. R. (2003). Task versus relationship conflict, team performance, and team member satisfaction: A meta-analysis. Journal of Applied Psychology, 88, 741-749.

DeRue, D. S., Hollenbeck, J., Ilgen, D., & Feltz, D. (2010). Efficacy dispersion in teams: Moving beyond agreement and aggregation. Personnel Psychology, 63, 1-40.

Dierdorff, E. C., & Ellington, J. K. (2013). Members matter in team training: Multilevel and longitudinal relationships between goal orientation, self-regulation, and team outcomes. Personnel Psychology.

Dierdorff, E. C., Bell, S. T., & Belohlav, J. B. (2011). The power of 'we': Effects of psychological collectivism on team performance over time. Journal of Applied Psychology, 96, 247-262.

Dominick, P. G., Reilly, R. R. & McGourty, J. W. (1997). The effects of peer feedback on team member behavior. Group & Organization Management, 22, 508-520.

Eby, L. T., & Dobbins, G. H. (1997). Collectivistic orientation in teams: An individual and group-level analysis. Journal of Organizational Behavior, 18, 275-295.

Erez, A., Lepine, J. A., & Elms, H. (2002). Effects of rotated leadership and peer evaluation on the functioning and effectiveness of self-managed teams: A quasi-experiment. Personnel Psychology, 55, 929-948.

Ferrell, O. C. & Gresham, L. G. (1985). A contingency framework for understanding ethical decision making in marketing. Journal of Marketing, 49, 87-96.

Gist, M. E. (1987). Self-efficacy: Implications for organizational behavior and human resource management. Academy of Management Review, 12, 472-485.

Gladstein, D. L. (1984). Groups in context: A model of task group effectiveness. Administrative Science Quarterly, 29, 499-517.

Gully S. M., Incalcaterra K. A., Joshi A., Beaubien J. M. (2002). A meta-analysis of team efficacy, potency, and performance: Interdependence and level of analysis as moderators of observed relationships. Journal of Applied Psychology, 87, 819–832.

Harkins, K., Harkins, S. & Latane, B. (1981). Identifiability as a deterrent to social loafing: two cheering experiments. Journal of Personality and Social Psychology, 40, 303-311.

Hinsz, V. B., Tindale, R. S., & Vollrath, D. A. (1997). The emerging conceptualization of groups as information processors. Psychological Bulletin, 121, 43-64.

Hooper, A.C., Cullen, M. J. and Sackett, P. R. (2006). Operational threats to the use of SJTs: faking, coaching, and retesting issues, in Weekley, J.A. and Ployhart, R.E. (Eds), Situational judgment tests: Theory, measurement and application. Mahwah, NJ: Erlbaum, pp. 205-32.

Hunt, S. D., & Vitell, S. (1986). A general theory of marketing ethics. Journal of Macromarketing, 6(1), 5-16.

Janis, I. L., & Mann, L. (1977). Decision making: A psychological analysis of conflict, choice, and commitment. London: Cassel & Collier Macmillan.

Janz, B. D., Colquitt, J. A., & Noe, R. A. (1997). Knowledge worker team effectiveness: The role of autonomy, interdependence, team development, and contextual support variables. Personnel Psychology, 50, 877-904.

Jassawalla, A., Sashittal, H., & Malshe, A. 2009. Students' perceptions of social loafing, its antecedents and consequences in undergraduate business classroom teams. Academy of Management Learning and Education, 8, 42–54.

Jehn, K. A. (1997). A qualitative analysis of conflict types and dimensions in organizational groups. Administrative Science Quarterly, 42, 530-557.

Johnson, C. E. (2007). Ethics in the workplace: Tools and tactics for organizational transformation.

Kaplan, R. S, & Norton, D. P. (1992). The Balanced Scorecard - Measures That Drive Performance. Harvard Business Review (January–February): 71–79.

Kaplan, R. S, & Norton, D. P. (1993). Putting the Balanced Scorecard to Work. Harvard Business Review.

Kaplan, R. S, & Norton, D. P. (1996). Using the Balanced Scorecard as a Strategic Management System. Harvard Business Review, 76.

Kaplan, R. S, & Norton, D. P. (1996). The Balanced Scorecard: Translating Strategy into Action. Boston, MA.: Harvard Business School Press.

Kaplan, R. S, & Norton, D. P. (2000). The Strategy-Focused Organization: How Balanced Scorecard Companies Thrive in the New Business Environment. Boston, MA: Harvard Business School Press.

Kirkman, B. L., Rosen, B., Tesluk, P. E., & Gibson, C. B. (2004). The Impact of team empowerment on virtual team performance: The moderating role of face-to-face interaction. Academy of Management Journal, 47, 175-192.

Kohlberg, L. A. (1984) The psychology of moral development: The nature and validity of moral stages (Vol 2.). San Francisco: Harper & Row

Kozlowski, S. W. J., & Bell, B. S. (2003). Work groups and teams in organizations. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), Handbook of psychology: Industrial and organizational psychology (Vol. 12, pp. 333-375). London: Wiley.

Latane, B., Williams, K. & Harkins, S. (1979). Many hands make light the work: the causes and consequences of social loafing. Journal of Personality and Social Psychology, 47, 822-832.

Lindsley, D. H., Brass, D. J., & Thomas, J. B. (1995). Efficacy-performance spirals: A multilevel perspective. Academy of Management Review, 20, 645-678.

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2000). A temporally based framework and taxonomy of team processes. Academy of Management Review, 26, 356-376.

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. Journal of Applied Psychology, 85, 273-283.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A. & Braverman, E. P. (2001). Predicting job performance using situational judgment tests: a clarification of the literature, Journal of Applied Psychology, 86, 730-40.

McDevitt, R., Giapponi, C., & Tromley, C. (2007). A model of ethical decision making: The integration of process and content. Journal of Business Ethics, 73(2), 219-229.

McIntyre, R. M., & Dickinson, T. L. (1992). Systemic assessment of teamwork processes in tactical environments (N61339-91-C-0145). Norfolk, VA: Old Dominion University.

McIntyre, R. M., & Salas, E. (1995). Measuring and managing for team performance: Emerging principles from complex environments. In R. A. Guzzo & E. Salas (Eds.), Team effectiveness and decision making in organizations (pp. 9-45). San Francisco: Jossey-Bass.

Mello, J. A. (1993). Improving individual member accountability in small work group settings. Journal of Management Education, 17, 253-259.

Morgan, B. B., Jr., Salas, E., & Glickman, A. S. (1993). An analysis of team evolution and maturation. The Journal of General Psychology, 120, 277-291.

Morgan, B. B., Jr., Salas, E., & Glickman, A. S. (1993). An analysis of team evolution and maturation. The Journal of General Psychology, 120, 277-291.

Nagle, B. (1953). Criterion development. Personnel Psychology, 6, 271-289.

Nguyen, N. T., Biderman, M. D., & McDaniel, M. A. (2005). Effects of response instructions on faking a situational judgment test. International Journal of Selection and Assessment, 13, 250–260.

Ohland, M. W. et al. (in press). The comprehensive assessment of team member effectiveness: Development of a behaviorally anchored rating scale for self and peer evaluation. Academy of Management Learning & Education.

Pearce, C. L., & Giacalone, R. A. (2003). Teams behaving badly: Factors associated with anticitizenship behavior in teams. Journal of Applied Social Psychology, 33, 58–75.

Prince, C., & Salas, E. (1993). Training and research for teamwork in the military aircrew. In E. L. Wiener, B. G. Kanki, & R. L. Helmreich (Eds.), Cockpit resource management (pp. 337-366). San Diego, CA: Academic Press.

Rest, M. (1985). Moral development; advances in research and theory. New York: Praeger.

Rousseau, V., Aube, C., & Savoie, A. (2006). Teamwork behaviors: A review and integration of frameworks. Small Group Research, 37, 540-570.

Salas, E., DiazGranados, D., Klein C., Burke, C. S., Stagl, K. C., Goodwin, G. F., & Halpin, S. H. (2008). Does team training improve team performance? A Meta-analysis. Human Factors, 50, 903-933.

Salas, E., Sims, D. E., & Burke, C. S. (2005). Is there a "big five" in teamwork? Small Group Research, 36, 555-599.

Salas, E., Sims, D. E., & Burke, C. S. (2005). Is there a "Big Five" in Teamwork? Small Group Research, 36, 555-599.

Salas, E., Stagl, K. C., Burke, C. S., & Goodwin, G. F. (2004). Fostering team effectiveness in organizations: Toward an integrative theoretical framework of team performance. In W. Spaulding & J. Flowers (Eds.), Modeling complex systems: Motivation, cognition and social processes. Lincoln: University of Nebraska Press.

Seers, A. (1989). Team-member exchange quality: A new construct for role-making research. Organizational Behavior and Human Decision Processes, 43, 118-135.

Seers, A., Petty, M. M., & Cashman, J. F. (1995). Team-member exchange under team and traditional management: A naturally occurring quasi-experiment. Group & Organization Management, 20, 18-38.

Smith-Jentsch, K. A., Johnston, J. H., & Payne, S. C. (1998). Measuring team-related expertise in complex environments. In J. A. Cannon-Bowers & E. Salas (Eds.), Making decisions under stress: Implications for individuals and team training (pp. 61-87). Washington, DC: American Psychological Association.

Spreitzer, G. M., Cohen, S. G., & Ledford, G. E. (1999). Developing effective self-managing work teams in service organizations. Group & Organization Management, 24, 340-366.

Stevens, M. J., & Campion, M. A. (1994). The knowledge, skill, and ability requirements for teamwork: Implications for human resource management. Journal of Management, 20, 503-530.

Stewart, G. L. (2006). A meta-analytic review of relationships between team design features and team performance. Journal of Management, 32, 29-54.

Taggar, S. & Brown, T. C. (2001). Problem solving team behaviors: Development and validation of BOS and a hierarchical factor structure. Small Group Research, 32, 698-726.

Tannenbaum, S. I., Beard, R. L., & Salas, E. (1992). Team building and its influence on team effectiveness: An examination of conceptual and empirical developments. In K. Kelley (Ed.), Issues, theory, and research in industrial/organizational psychology (pp. 117-153). New York: Elsevier Science.

Tesluk, P. E., & Mathieu, J. E. (1999). Overcoming roadblocks to effectiveness: Incorporating management of performance barriers into models of work group effectiveness. Journal of Applied Psychology, 84, 200-217.

Trevino, L. K. (1986). Ethical decision making in organizations: A person-situation interactionist model. Academy of management Review, 601-617.

Trevino, L. K., & Nelson, K. A. (2010). Managing business ethics. Wiley.

Webber, S. S. (2002). Leadership and trust facilitating cross-functional team success. Journal of Management Development, 21, 201-214.

Weldon, E., & Weingart, L. R. (1993). Group goals and group performance. British Journal of Social Psychology, 32, 307-334.

Weldon, E., Jehn, K. A., & Pradhan, P. (1991). Processes that mediate the relationship between a group goal and improved group performance. Journal of Personality and Social Psychology, 61, 555-569.

Wilhelm, W. J. (2006). Ethical reasoning for the business classroom: A decision-making framework. Journal of Applied Research for Business Education, 4(2), 1-7.

Zaccaro, S. J., Rittman, A. L., & Marks, M. A. (2001). Team leadership. Leadership Quarterly, 12, 451-483.

ADDENDUM

Overview

This addendum describes the results of additional analysis of Capsim's Comp-XM®. For more comprehensive technical details regarding Comp-XM®, readers are encouraged to consult the full technical synopsis. In brief, Comp-XM is a simulation-based assessment tool designed to broadly capture "business acumen" – knowledge and skills related to the various functions of organizational management. Comp-XM goes beyond measuring recall of acquired knowledge assessed by standardized exams in a static testing environment. Comp-XM instead provides both an assessment of acquired knowledge and an assessment of how effectively individuals use this knowledge to manage a business in a changing and competitive simulated market.

The paragraphs below describe the procedure and results that were followed in factor analyzing Comp-XM data.

Procedure

Sample data on the 41 default Comp-XM "board query" (BQ) items comprised 26,166 individuals that completed the assessment between January 2013 and April 2014. These individuals were students at the undergraduate and graduate levels from a variety of post-secondary educational institutions.

These data were subjected to exploratory factor analysis; more specifically, principal component analysis (PCA). Because the data were dichotomous (correct/incorrect responses), polychoric correlations were first computed using a SPSS routine developed by Lorenzo-Seva and Ferrando (2014). The use of polychoric correlations, versus Pearson correlations, is advisable when data are not normally distributed or have extreme response distributions (Lee, Zhang, & Edwards, 2012).

The polychoric correlation matrix generated from the sample Comp-XM data was the input for a subsequent PCA using varimax rotation. The factor number retention was initially examined using the Kaiser criterion (eigenvalues greater than 1.0); however, interpretability of the factor structure was emphasized in the final determination of the number of factors to rotate.

Results

PCA results suggested a five-factor solution based on the Kaiser criterion. This solution accounted for 49% of the cumulative variance. Upon inspection of the factor loadings, however, only two of the 41 items had loading greater than .30 on the fifth factor. Thus, a four-factor solution was specified. The four-factor solution accounted for 46% of the cumulative variance. The following table displays the factor loadings from this PCA. Item labels are provided and the items are sorted into conceptual categories (e.g., finance/accounting) to facilitate interpretation.

Rotated Factor Loadings from PCA

Labels	Factor 1	Factor 2	Factor 3	Factor 4
A-16: Break Even Analysis	0.243	0.125	0.133	0.514
A-279: Understanding the Accounting Equation	0.298	0.448	0.298	0.106
A-286: Revenue Recognition	0.474	0.261	-0.037	0.194
A-290: Identifying Fixed vs. Variable Costs	0.359	0.557	0.376	0.153
A-320: Calculating Book Value	0.622	0.206	0.119	0.109
A-362: Identifying Change in Equity	0.203	0.540	0.259	0.201
A-364: Interpreting the Cash Flow Statement	0.390	0.486	0.254	0.15
A-389: Understanding Item Carrying Values on the Balance Sheet	0.143	0.473	0.318	0.105
A-94: Understanding the Accounting Equation	0.500	0.336	0.215	0.245
F-115: DuPont Analysis	0.163	0.567	0.168	0.24
F-260: Calculating Dividend Yield	0.599	0.193	0.225	0.049
F-261: Effects of Change in Depreciation Expense on Financial Statements	0.141	0.118	0.106	0.517
F-267: Calculating Simple Ratios	0.257	0.570	0.315	0.055
F-274: Calculating Ratios from the Annual Report	0.147	0.569	0.349	0.129
F-301: Calculating Stock Repurchase	0.589	0.317	0.305	0.044
F-347: Cash Management	0.478	0.194	0.251	0.18
F-358: Effect of Investment Decisions	0.382	0.708	0.297	0.157
F-41: Retirement of Debt	0.055	0.212	0.034	0.66
H-319: Calculating Productivity Impact	0.287	0.242	0.623	0.144
H-323: Calculating Recruiting Costs	0.076	0.295	0.349	0.13
H-324: Calculating Training Costs	0.521	0.431	0.233	-0.025
H-326: Calculating Separation Costs	0.607	0.179	0.125	0.226
H-329: Calculating Future Labor Wages	0.338	0.335	0.413	0.137
O-351: TQM Break Even Analysis	0.203	0.162	0.658	0.22
M-308: Identifying Marketing Efficiency	0.114	0.187	0.482	0.308
M-334: Forecasting	0.314	0.182	0.595	0.07
M-343: Creating Marketing Budgets	0.096	0.219	0.621	0.268
M-346: Identifying Competitors Using the "Four P's"	0.067	0.292	0.49	0.079
M-349: Demand Analysis	0.309	0.307	0.598	0.057
M-354: Identifying Price Elasticity	0.285	0.234	0.601	0.173
M-355: Market Sizing	-0.041	0.183	0.557	-0.008
O-322: Operational Impact of Unit Margin	0.476	0.328	0.143	0.133
P-258: Capacity Analysis	0.564	-0.055	0.35	0.396
P-309: Determining Acceptable Inventory Levels	0.382	0.196	0.186	0.274
P-328: Cost of Right-Sizing Plant	0.437	0.231	0.258	0.313
P-682: Plant Utilization	0.705	0.142	0.25	0.195
S-363: Identifying Tactics for Building Competitive Advantage	0.386	0.812	0.291	0.23
S-668: Developing Mission/Vision Statements	0.345	0.384	0.349	0.056
S-671: Strategic Analysis	0.293	0.441	0.35	0.035
S-674: Competitive Analysis	0.514	0.269	0.264	0.128
S-73: Identifying Strategies	0.364	0.156	0.434	-0.087

References

Lee, C. T., Zhang, G., & Edwards, M. C. (2012). Ordinary least squares estimation of parameters in exploratory factor analysis with ordinal data. Multivariate Behavioral Research, 47, 314–339.

Lorenzo-Seva, U., & Ferrando, P. J. (2014). POLYMAT-C: A Comprehensive SPSS program for computing the polychoric correlation matrix. *Behavior Research Methods*. DOI 10.3758/s13428-014-0511-x.