Improving Survey Efficiency: Understanding the Relationships among Common Metrics for Concept Evaluation

Jennifer G. Boldry, Michael Polster and Susan McDonald

In order to minimize risks to data quality associated with respondent burden and fatigue, the market research industry has become increasingly concerned about balancing survey scope/value and length. Inspired by these concerns, a commercial market research company has applied meta-analytic techniques to provide empirical guidance in improving the efficiency of concept evaluation research. In the present research, we meta-analytically examine relationships among metrics that are commonly included in concept testing research (i.e. ratings of how "compelling", "credible", and "unique" the concept is) in order to understand whether these concepts (a) measure independent constructs and (b) predict the likelihood of a customer to act (i.e. seek additional information or purchase a product/service). Results based on the data from 10 independent studies across three industries (Healthcare, Information Technology, and Transportation) suggest that (1) ratings of "compelling," "credible", and "unique" are correlated with one another, and thus that (2) use of all three metrics as predictors of "likely to take action" nets little incremental gain in predictive validity over the most correlated metric as a single predictor (i.e., compelling). Taken together, these findings indicate that it is possible to limit the number of dimensions on which new concepts are evaluated without sacrificing significant decision-making support.

Keywords: Survey efficiency, concept evaluation

Introduction

Testing the appeal of new concepts (e.g., products, services, advertisements) is a common and crucial survey mandate in the market research industry. The primary goal of concept evaluation research is often to predict the likelihood that members of a target audience will take action (e.g. seek more information; purchase a new product/service). Indeed, direct ratings of the likelihood to take action are commonly included in concept evaluation research as the key endpoint on which decisions are made. However, additional metrics (e.g., compelling, unique, credible) are also often included for one of three reasons: (1) they are believed to improve precision by measuring the same underlying construct; (2) they are believed to measure independent constructs; or (3) it is inappropriate to include the "likelihood to take action" metric and the additional measures are being used as surrogates for that endpoint. There has, however, been little research examining the relationships among common concept testing metrics, leaving analysts to make survey design and data analysis decisions with little empirical guidance regarding which measures (if any) should be prioritized. The present research was designed to provide empirical guidance regarding the value of including multiple metrics in concept evaluation research.

Theoretical Approaches to Survey Development

Nearly three decades ago, Churchill (1979) and Peter (1979) implored the market research industry to pay more attention to measurement issues in survey development. Since then,

approaches to survey construction based on classical test theory have dominated. Classical test theory suggests that "true scores" are a function of observed scores plus error (e.g., Nunally & Bernstein, 1994). Consequently, true scores can be most accurately estimated when multiple observed measures are collected because the common variance among measures can be treated as the "true score" and error (measurement or random) can be partialled out. In other words, a particular construct can be measured most accurately through inclusion of multiple measures of the construct. In this tradition, measurement reliability (precision) is typically assessed using the average inter-item correlation (coefficient α). Higher average inter-item correlations indicate more reliable measures. That is, the more highly the items are correlated, the more likely that they are measuring the same underlying construct. It is the classical test theory tradition that motivates researchers to include metrics other than "likely to take action" in an effort to improve the precision of their measurement metrics in concept evaluation research. In fact, Churchill (1979) suggested that "Marketers are much better served with multi-item then single-item measures of their constructs…" (p. 66).

Of course, the classical test theory approach poses a practical challenge in the market research industry, where there is constant pressure on researchers to reduce respondent burden and survey length while simultaneously maximizing content coverage in each survey. In fact, including multiple measures of a construct not only limits the number of concepts that can be examined, but can compromise data quality as respondents succumb to fatigue or become frustrated with completing multiple versions of seemingly equivalent items (Drolet & Morrison, 2001). Consequently, applied market researchers are left with the conundrum of choosing between item precision (reliability) and cost (in dollars, survey "bandwidth", and data quality).

Not surprisingly, there have been challenges to classical test theory approaches to survey development. In fact, it has been suggested that single-item measures are sufficient for measuring psychological constructs if the construct of interest is concrete or narrowly defined (e.g., Rossiter, 2002; Sackett & Larson, 1990). For example, Scarpello and Campbell (1983) concluded that a single-item measure was preferable to a scale measure of overall job satisfaction. Similarly, Wanous et al. (1997) meta-analytically compared single-item measures against scale measures of job satisfaction and found that single-item measures converged substantially with scale measures (corrected $r = .67$).

As an alternative approach to scale development, researchers sometimes include metrics believed to be predictive of outcomes of interest as proxy metrics when including direct measures of the outcomes of interest is inappropriate. In psychometric terms, researchers in this case are applying standards of predictive validity to select metrics for inclusion. *Predictive validity* reflects whether a construct of interest is related to some outcome of interest and is also typically evaluated by examining the association between a measure of interest and an outcome of interest (e.g., via regression). For example, the SAT exhibits predictive validity if SAT scores predict academic performance in college.

The primary goal of the present study is to generate empirical evidence that would provide guidance to those seeking to balance concerns about reliability (precision) and predictive validity against limitations in survey bandwidth in concept evaluation research. To meet these goals we employ two methods. First, to evaluate whether metrics commonly included in concept

evaluation research are independent or multiple measures of the same underlying construct, we meta-analytically examine relationships among ratings of compelling, credible, and unique in the context of new products/concepts. Second, we explore the predictive validity of these measures with respect to the likelihood of taking action.

Method

Criteria for Inclusion

We included studies conducted by a commercial market research company that provided measures of at least two of the following metrics on quantitative scales: compelling, credible, unique, likely to take action. A total of 10 independent studies, based on responses from 2,140 participants across three industries (IT, Transportation, Healthcare) qualified for inclusion.

Calculation of Effect Sizes

Up to six Pearson's correlations were computed per study: compelling/credible, compelling/unique, credible/unique, compelling/likely, credible/likely, unique/likely[1].

When approaching a meta-analysis, the researcher has a choice between two statistical models: random-effects or fixed-effects (Field, 2001; Hedges & Vevea, 1998; Hunter & Schmidt, 2000). Fixed-effects models are more powerful when their homogeneity assumption (i.e., all effect sizes estimate a common population effect) is satisfied. However, when the latter assumption is not satisfied, fixed-effect models underestimate standard errors of parameter estimates and inflate the Type I error rate (i.e., underestimate confident intervals). Monte Carlo simulations, for example, suggest that the Type I error rate in heterogeneous fixed-effects models ranges between .43 and .80, which is dramatically higher than the nominal .05 level (Field, 2003). In all of the analyses reported below, we initially tested the homogeneity assumption of the fixed-effects model (which is equivalent to the test of random-effects variance). The homogeneity assumption was met in all instances so we employed fixed-effects models. To determine whether aggregated effect sizes differed between moderator categories, we used the χ^2 distributed *QB* statistic[2].

Four of the 10 studies that met the inclusion criteria measured more than one product/concept within the same respondents. Treating each product/concept as though it were evaluated in an independent study would effectively weight studies that measured perceptions of multiple products/concepts more heavily than those that measured perceptions of a single produce/concept. Consequently, for studies that measures perceptions on multiple concepts, we averaged the correlations within each study such that each independent group of respondents contributed only one correlation to the analysis.

[1] Although there is a well-documented slight downward bias in *r* as an estimate of the population correlation, Hunter and Schmidt (1990) recommend analyzing unadjusted Pearson's correlations because the bias is less than rounding error when sample sizes are 40 or larger and, more importantly, because Fisher's z transformation leads to potentially substantial bias in the opposite direction independent of sample size.

[2] QB comparisons were also computed based on Cohen's d effect sizes rather than Pearson's r effect sizes. Results were comparable so the more parsimonious analysis is presented here.

Results

Correlations between Measures

Table 1 displays the pairwise correlations between the three dimensions overall and for each industry. Overall, there is strong correlation between the compelling and credible dimensions (r = .50), but the size of that correlation differs across the three industries (QB (2) = 17.00, $p < .01$). The correlation between compelling and credible is statistically significant for all industries, but the relationship is significantly weaker in the IT industry than in either the Transportation or Healthcare industries based on pairwise comparisons with Scheffe adjustments. There is also a moderate-to-strong overall correlation between compelling and unique (r = .43). Once again, however, the size of that correlation differs across industries QB (2) = 34.45, $p < .01$ (see Table 1), though is statistically significant for all three industries. Finally, there is a moderate correlation overall between credible and unique (r = .34), which again, differs in magnitude across industries (QB (2) = 48.96, $p < .01$). In this instance, the correlation between credible and unique is not statistically significant in the IT industry, and significantly weaker in the IT industry than in either the Transportation or Healthcare industries (where the correlations are both statistically significant) based on pairwise comparisons with Scheffe adjustments.

Table 1. Correlations among the Primary Measures

Measure	Industry	k[a]	r	95% CI	QB[b]
Compelling-Credible	***Overall***		***.50***	***.44/.56***	**17.00****
	IT	6	.31	.21/.42	
	Transportation	2	.61	.42/.80	
	Healthcare	2	.58	.50/.66	
Compelling-Unique	***Overall***		***.43***	***.37/.49***	**34.45****
	IT	6	.17	.07/.28	
	Transportation	2	.53	.34/.72	
	Healthcare	2	.56	.48/.64	
Credible-Unique	***Overall***		***.34***	***.28/.40***	**48.96****
	IT	6	.03	-.08/.13	
	Transportation	2	.46	.27/.64	
	Healthcare	2	.49	.41/.57	

[a]Indicates the number of effect sizes.
[b]*QB* tests for differences across industries with df equal to one less than the number of industries (e.g., df = 2).
**$p < .01$

Correlations with Likelihood to Take Action

Table 2 shows the correlations between each dimension and the likelihood of taking action (overall and by industry). There is a strong overall correlation between compelling and likely to take action (r = .60) that differs in size by industry (QB (2) = 7.72, $p < .05$). There is a moderate

to strong correlation between credible and likely to take action (r = .44). However, comparison of the correlation across the three industries indicates that the correlation differed across industry QB (2) = 25.70, $p < .01$. Finally, there is a moderate-to-strong correlation between unique and likely to take action (r = .40), which is again different in size for the three industries (QB (2) = 23.55, $p < .01$).

Table 2. Correlations between the Primary Measures and Likely to Take Action

Measure	Industry	k[a]	r	95% CI	QB[b]
Compelling-Likely	***Overall***		***.60***	***.52/.69***	**7.72***
	IT	4	.47	.33/.60	
	Transportation	2	.72	.53/.91	
	Healthcare	1	.71	.56/.86	
Credible-Likely	***Overall***		***.44***	***.35/.52***	**25.70****
	IT	4	.20	.07/.33	
	Transportation	2	.52	.33/.71	
	Healthcare	1	.69	.55/.84	
Unique-Likely	***Overall***		***.40***	***.32/.49***	**23.55****
	IT	4	.17	.04/.30	
	Transportation	2	.46	.27/.65	
	Healthcare	1	.65	.51/.80	

[a]Indicates the number of effect sizes.
[b]QB tests for differences across industries with df equal to one less than the number of industries (e.g., df = 2).
*$p < .05$
**$p < .01$

Predicting the Likelihood to Take Action

To investigate the predictive validity of the primary measures relative to the likely to take action metric, the following regression equations were computed: (1) "compelling" to predict likely to take action, (2) "compelling" and "credible" to predict likely to take action, (3) compelling, credible, and unique to predict likely to take action. From each of these equations, the percentage of variance accounted for (r^2) was averaged across studies and compared (see Table 3). Overall, results indicate little incremental gain in model fit when credible is entered as a second predictor to compelling (compelling alone: r^2= 37% vs. compelling and credible: r^2= 42%). Adding unique as a third predictor adds even less (r^2= 43% when all three variables are used). Notably, the pattern is similar across all three industries. A model that included pairwise and a three-way interaction term failed to offer a better fit.

Table 3. Percentage of Variance Accounted For in Predicting Likely to Take Action

Industry	Regression	k[a]	r^2
Overall			
	Compelling		37%
	Compelling and Credible		42%
	Compelling, Credible, Unique		43%
IT		4	
	Compelling		22%
	Compelling and Credible		24%
	Compelling, Credible, Unique		25%
Transportation		2	
	Compelling		52%
	Compelling and Credible		53%
	Compelling, Credible, Unique		54%
Healthcare		1	
	Compelling		50%
	Compelling and Credible		56%
	Compelling, Credible, Unique		58%

[a]Indicates the number of studies.

To examine the classical test theory tenet that there can be value in treating primary measures as indicators of a single underlying construct, primary measures were combined and a series of regression equations were computed. Specifically, for each study that included the primary measures and likely to take action, the following regression equations were computed (1) mean of compelling and credible to predict likely to take action and (2) mean of compelling, credible, and unique to predict likely to take action. From each of these equations, the percentage of variance accounted for (r^2) was averaged across studies and compared (see Table 4). Once again, results indicate that there is little incremental gain in model fit when composite measures are used to predict likely to take action, and the pattern of results is similar across industries.

Discussion

These findings provide empirical evidence regarding the value of including multiple metrics in concept evaluation research across three industries (Healthcare, IT, and Transportation), and suggest that there is little, if any, utility to collecting ratings on multiple dimensions beyond the likelihood to act on a proposition. Specifically, ratings of new concepts on traditional metrics like compelling, credible and unique are generally correlated, and in the case of compelling and credible, highly correlated across all three industries. Furthermore, using all three dimensions to predict likelihood to act nets little incremental gain in the percentage of variance accounted for over simply using ratings of how compelling an item is. Taken together, these findings suggest that it is possible to limit the number of dimensions on which new concepts are evaluated without sacrificing significant decision-making support.

Table 4. Percentage of Variance Accounted for in Predicting Likely to Take Action

Industry	Regression	k[a]	r^2
Overall			
	Mean of Compelling and Credible		38%
	Mean of Compelling, Credible, Unique		38%
IT		4	
	Mean of Compelling and Credible		19%
	Mean of Compelling, Credible, Unique		19%
Transportation		2	
	Mean of Compelling and Credible		49%
	Mean of Compelling, Credible, Unique		48%
Healthcare		1	
	Mean of Compelling and Credible		56%
	Mean of Compelling, Credible, Unique		56%

[a]Indicates the number of studies.

In addition to demonstrating that these additional dimensions offer limited utility when considered individually, these findings also indicate that they offer no additional predictive validity as a multiple-item measure as classical test theory might suggest. Indeed, a computed score of the average ratings of a concept on multiple dimensions did not predict likelihood of taking action any better than the compelling rating used in isolation. These results are consistent with previous findings that single-item measures are as reliable as multiple-item measures in some contexts (Gardner, Cummings, Dunham, & Pierce, 1998; Wanous, Reichers, & Hudy, 1997).

The present analysis could not, of course, comment on the validity of the likely to act construct itself. In fact, there is a vast literature suggesting that self-reported intentions are not perfectly correlated with actual behavior, leaving the door open to the possibility that additional metrics could have alternative diagnostic utility. Nor could the present analysis address whether metrics such as compelling, credible, and unique have diagnostic utility beyond predicting the likelihood to take action metric. However, given the correlations among these metrics it is questionable whether these metrics would be diagnostic in useful ways beyond predicting the likelihood of taking action. Finally, the present analysis focuses on a non-conjoint approach to concept testing research, and as such does not rule out the likelihood that multiple-item measures may be appropriate in other contexts or may serve to meet alternate research goals. For example, perceptions of concept uniqueness may not provide much in the way of additional utility when the research goal is to predict the likelihood of taking action, but may provide valuable insight into developing an effective marketing strategy for a new concept or product. In the final analysis, appropriately balancing research goals against the cost of including additional measures always lies within the purview of the researcher.

It is notable that the relative strength of the various relationships held across industries, even though the relationships among the constructs are weaker in the IT industry than in other industries. Industry differences are potentially attributable to several factors. First, the

differences could reflect variability in how constructs were operationally defined across industries (it was necessary to use different wording because the nature of the new concepts differed by industry). We have no clear avenue for addressing this possibility within the context of the present research, but do note that similarities in correlations between dimensions with Healthcare and Transportation despite use of different wording, is suggestive that other factors are involved. A second possibility is that differences in the scales used across the industries might account for the variability in findings across industries. Specifically, Healthcare studies used 10-point scales, whereas IT and Transportation studies used 5-point scales. This possibility therefore, also seems unlikely because results were more similar between Healthcare and Transportation, which used different scales, than between Transportation and IT, which used the same scales. A third, more viable explanation is that participants in the various industries evaluate concepts differently because they are more or less abstract. For example, in the IT industry, the concepts were largely IT services, which can be difficult to explain and visualize. By contrast, in the Transportation and Healthcare industries the concepts represented concrete products. This difference in the level of concept abstraction appears to provide the most plausible explanation for differences observed across industries, but certainly requires further investigation, which could occur most directly by systematically varying the nature of the concepts tested within the various industries.

References

Churchill, G.A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research, 6*, 64-73.

Drolet, A.L. & Morrison, D.G. (2001). Do we really need multiple-item measures is services research? *Journal of Services Research, 3*, 196-204.

Field, A.P. (2003). The problems using fixed-effects models of meta-analysis on real-world data. *Understanding Statistics, 2*, 77-96.

Field, A.P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods, 6*, 161-180.

Gardner, D.G., Cummings, L.L., Dunham, R.B., & Pierce, J.L. (1998). Single-item versus multiple-item measurement scales: An empirical comparison. *Educational and Psychological Measurement, 58*, 898-915.

Hedges, L.V. & Vevea, J.L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods, 3*, 486-504.

Hunter, J. E. & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative knowledge in psychology. International Journal of Selection and Assessment, 8, 275-292.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Newbury Park: Sage Publications.

Nunally, J.C. & Bernstein, I.H. (1994). *Psychometric Theory (3rd edition)*. New York: McGraw Hill.

Peter, J.P. (1979). Reliability: A review of psychometric basics and recent marketing practices. *Journal of Marketing Research, 6*, 6-17.

Rossiter, J.R. (2002). The C-OAR-SE procedure for scale development in marketing. *International Journal of Research in Marketing, 19*, 305-335.

Sackett, P.R. & Larson, J.R. Jr. (1990). Research strategies and tactics in industrial and

organizational psychology. In M.D. Dunnette & L.M. Hough (Eds.), *Handbook of Industrial and Organizational Psychology* (2nd ed.), vol. 1, 419-489. Palo Alto, CA: Consulting Psychologists Press.

Scarpello, V. & Campbell J.P. (1983). Job satisfaction: Are all the parts there? *Personnel Psychology*, *36*, 577-600.

Wanous, J.P., Reichers, A.E., & Hudy, M.J. (1997). Overall job satisfaction: how good are single-item measures? *Journal of Applied Psychology*, *82*, 247-252.

Jennifer G. Boldry is a methods specialist, Michael Polster is a vice-president, and Susan McDonald is the CEO of National Analysts Worldwide, 1835 Market St. 25th Floor, Philadelphia, PA 19103. **jboldry@nationalanalysts.com**.