Original Article

# Measuring consumer-based brand equity across brand portfolios: Many-facet Item Response Theory perspective

Received (in revised form): 8th October 2012

**Luming Wang**
completed her PhD at the University of Alberta (Canada) in 2010. Her research interests include branding, brand equity and brand management, research methodology, Bayesian modeling, consumer choice modeling, and international marketing.

**Adam Finn**
completed his PhD at the University of Illinois (Urbana, USA) in 1983. His research interests include marketing measurement, retailing and service assessment, new product management, retailing and service management, application of choice modeling to policy issues in marketing and the cultural industries.

**ABSTRACT** **Wang and Finn (in press) used Multivariate Generalizability Theory (MGT) to develop the first brand level consumer-based brand equity (CBBE) scale. It is an important addition to CBBE measurement literature. However, in general, the goal of MGT is to examine measurement scale in terms of reliability and to deal with the precision of observed raw data. It has little interest in correcting the observed measures (that is, brand ratings) for differences among the various elements encountered by each brand during a data collection design. The elements such as individual differences, dimensions and items are all confounded with CBBE measures. In order to address this issue, we introduce Many-facet Item Response Theory to complement the information provided by MGT analysis. These two measurement approaches, used together, allow for balanced and thorough interpretations of brand performance assessment data and offer better ways of improving performance measurement.**
*Journal of Targeting, Measurement and Analysis for Marketing* (2012) **20,** 254–260. doi:10.1057/jt.2012.20; published online 12 November 2012

**Keywords:** Many-facet Item Response Theory; Bayesian analysis; consumer-based brand equity; brand portfolio

## INTRODUCTION

Consumer-based brand equity (CBBE) is defined as the added value that a brand name endows upon a product as a result of a firm's marketing efforts, measured from the psychology approach. It has been considered as a sustainable competitive advantage that positively affects companies' future profits and long-term cash flow.[1] Numerous measures of CBBE[2,3] have been used since it was first conceptualized by Aaker[4] and Keller.[5] Aaker[4] identified five components of brand equity as awareness, associations, perceived quality, loyalty and other proprietary assets such as patents and trademarks. Keller[5] focused on brand knowledge, including awareness and unique favorable beliefs. More CBBE components such as willing to pay a price premium and uniqueness were later identified

**Correspondence:** Luming Wang
University of Manitoba, 676 Drake Center, Winnipeg R3T 5V4, Canada

 

by CBBE researchers.[3] Most of them are developed to differentiate consumers from each other.

Wang and Finn[6] used Multivariate Generalizability Theory (MGT) to develop the first brand level CBBE scale. It was designed to maximally differentiate brands from each other and reliably capture their differences in a brand portfolio context. MGT approach quantifies all the design factors involved and increases the validity and reproducibility of the CBBE estimates. But it has no interest in correcting the observed measures (that is, brand ratings) for differences among the various elements (such as individual respondents, dimensions, items) encountered by each brand during a data collection design. In order to achieve a contextually unbiased, objective estimate of brand equity, individual differences, together with other extraneous sources, must be accounted for. Removing the biases can also make the outcome of the measurement to be generalized beyond the specific elements of the measurement, such as the particular consumers, dimensions, items and so on. To address this issue, we introduce Many-facet Item Response Theory (MFIRT) to complement the information provided by MGT analysis.

In general, the goal of MGT is to examine measurement scale in terms of some index of reliability, deal with the precision of observed raw data and optimize a future data collection design by minimizing the influence of identified sources of error variances, whereas the purpose of MFIRT is to correct the observed measure for extraneous differences within each facet of a data collection design.[7] These two approaches are complementary. MGT approach provides general, group-level information particularly in making overall decisions about questionnaire design and the unit of analysis is the total score, whereas MFIRT offers more specific information at the individual respondent, item, dimension and interaction level and the unit of analysis is individual brand rating. These two measurement approaches, used together, allow for balanced and thorough interpretations of brand performance assessment data and offer better ways of improving performance measurement.

The current research focuses on how to apply MFIRT to get objective and unbiased estimates of brands in a brand portfolio context (including both master brands such as Coca-Cola and sub-brands such as Diet Coke and Coke Zero) and how sub-brands differentiate themselves from each other and from the master brands in terms of the brand equity value. As in Wang and Finn's paper (in press), brand portfolio in the current work is defined as a branded house rather than a house of brands.[8] A branded house consists of a single master brand and its sub-brands. For example, Disney is the master brand and Disney World, Disney Land, Disney Picture and Disney DVD can be seen as sub-brands. In contrast, a house of brands contains independent, unconnected brands. Consumers may not realize that the independent and unconnected house of brands belong to the same company's portfolio. For example, when consumers buy Folgers Coffee, most are probably unaware that Folgers is a division of Procter & Gamble.

## METHODOLOGY

Item Response Theory (IRT) is an advanced testing theory that has been accepted as a better alternative to Classical Test Theory (CTT) in the areas of education psychology and psychometrics.[9,10] It is the simplest two-facet (that is, consumers and items) version of the more general MFIRT. The marketing literature has relied largely on the application of CTT-based approaches for long time (see Balasubramaniam and Kamakura[11] and Singh *et al*[12] as exceptions). More recently, more applications of IRT in the marketing literature have started to appear.[13–15] But these are still traditional IRT modeling in terms of facets. For example, de Jong *et al*[13] examined measurement invariance in a cross-country study using an IRT model. In their study, consumers were nested within countries and country was treated as an extraneous factor rather than a research focus. There was no separate country facet in their model.

Traditional IRT models cannot fully capture the complexity of CBBE (that is, multi-dimensional with multiple sources of systematic variance) especially in a brand portfolio context. We introduce MFIRT, originally developed by Linacre[16] in educational psychology, which extends

IRT models from two facets (that is, consumers and items), to model additional facets (such as brands and dimensions) in the responses. Moreover, we combine MFIRT method with the Rating scale model[17] to accommodate both the multi-facet and the ordinal character of CBBE data.

In order to estimate the brand equity of master brands and sub-brands simultaneously after accounting for extraneous sources of variance (for example, individual differences, dimensions, items and response categories), we specify the model as follows:

$$\log(P_{rbmdik} / P_{rbmdi(k-1)}) = B_b + M_m - P_r - I_i - D_d - C_k \quad (1)$$

where $P_{rbmdik}$ and $P_{rbmdi(k-1)}$ are the probability of choosing category $k$ and $k-1$, respectively, on item $i$ in dimension $d$ for person $r$ on sub-brand $b$ which belongs to master brand $m$. $P_r$ is the criticality of person $r$; $B_b$ is the brand equity of sub-brand $b$; $M_m$ is the brand equity of master brand $m$; $I_i$ is the threshold of item $i$; $D_d$ is the dimensional threshold of dimension $d$; $C_k$ is the threshold of category $k$ relative to category $k-1$.

$B_b$ and $M_m$ are the major research interest. After accounting for extraneous factors (such as $P_r$, $I_i$ and $D_d$) that are confounded with $B_b$ and $M_m$ in the brand equity evaluations, $B_b$ and $M_m$ give us contextually unbiased estimates. Individual differences are examined by $P_r$. A high threshold for $P_r$ means that this particular consumer tends to give lower responses to all the brands evaluated, other things (including brand equity) being equal. It could be interpreted that this consumer is harder to please. By introducing $C_k$, the proposed approach makes a weaker ordinal assumption about the categories.[18] In particular, ordered step calibration is designed to capture the phenomenon that as a consumer moves up the leniency continuum or as a brand moves up the brand equity continuum, each response category in turn becomes the most probable response. Evidence of disordered average measurement and/or step calibration questions the functioning of the Likert scale. We use $I_i$ and $D_d$ to examine the different characteristics of items nested within dimensions. A high threshold for $I_i$ or $D_d$ means it is harder for consumers to agree with a specific item/dimension. A low threshold means an item/dimension is agreed to more easily.

Moreover, in the model, (a) the contribution of each facet (that is, consumers, brands, items nested within dimensions and response categories) to the evaluations is captured in a single parameter with a value independent of all other parameters within the frame of reference. It is the brand parameter value (that is, a brand's equity) that a CBBE measurement study is majorly intended to determine. This single parameterization is necessary if, for example, brands are to be arranged in one order of merit. (b) The parameters combine additively to produce the evaluations. Additivity implies that all the parameters share one linear scale. Not only does linearity assist brand managers in understanding the meaning underlying the data, it also provides a useful basis for further analysis, because many commonly used statistical procedures assume linearity in their data. (c) The estimate of any parameter is dependent on the accumulation of all responses in which it participates, but is independent of the particular value of any of those responses. This axiom of 'local independence' allows the statistical estimates of the measures to be as free as possible of which particular consumers evaluate which particular brands on which particular dimensions and items.[19] Therefore, the proposed approach not only has meaning generalized beyond the local details of the evaluation situation,[20] but also has the capability to accommodate more flexible data collection designs and handle missing data.

## EMPIRICAL APPLICATION

The fully crossed (brands by respondents) data set used by Wang and Finn[6] is examined here to demonstrate the complementary roles played by MFIRT to MGT analysis. It was collected in 2008 from 254 North American undergraduate students, who received course credit for their participation. The students evaluated six soft drink brands (that is, Classic Coke, Diet Coke,

Coke Zero, Pepsi, Diet Pepsi and Diet Pepsi Max). The brand evaluation instrument consisted of 25 items (as in Table 3) capturing the five dimensions (that is, brand awareness, brand associations, brand loyalty, perceived quality and uniqueness). All items were positively worded and used a 7-point Likert scale from 'strongly disagree' (1) to 'strongly agree' (7).

We choose a Bayesian approach and use WinBUGs to estimate the brand portfolio model. The total brand equity of a sub-brand is modeled as the sum of its individual brand equity and the equity it gets from its master brand.

Bayesian analysis requires the specification of the models based on prior distributions for model parameters and hence random samples are simulated from the posterior distributions through simulated Markov chain procedures.[21] To assess the convergence and Monte Carlo variances, we use two chains with differing starting values and calculate the Gelman–Rubin statistic, as modified by Brooks and Gelman,[22] which compares within- to between-chain variability. The ratio $R = B/W < 1.05$ for all parameter estimates signifies that convergence is attained. A total of 12 000 iterations were obtained with the first 4000 as burn-in.

Model identification is attained by imposing necessary constraints. The vectors of brand equity dimension thresholds ($D_d$) sum to zero and the main effect for master brand Coca-Cola is fixed at 1. The sub-brand main effects ($B_b$) are also set to sum to zero within each family brand group. Similarly, item thresholds ($I_i$) are set to sum to zero within each dimension. Consumer criticality ($P_r$) is drawn from a normal distribution with the mean fixed to zero.

In Table 1, we list estimates of CBBE for all six sub-brands with the two master brands after accounting for the influence of persons, items and dimensions.

When comparing the two master brands, the Coke family has higher CBBE than the Pepsi family (1 versus 0.807). After accounting for the master brands, Classic Coke as a sub-brand itself has the highest CBBE (0.266) and Diet Coke has the lowest one (−0.187). After combining with the brand equity sub-brands get from their master brands, Classic Coke still has the highest brand equity in total (1.266), followed by Original Pepsi (1.036). The sub-brand with the lowest combined brand equity is Diet Pepsi Max (0.634). From the results, we can see that the sub-brands within a brand portfolio do not have the same brand equity. In our empirical examination of CBBE, the sub-brands are different from their master brands and from each other.

Estimates of respondents' criticality ($P_r$), item threshold ($I_i$) and dimension threshold ($D_d$) are also shown in Tables 2–4.

The mean criticality of consumers is −0.005, with a standard deviation of 0.283. The range is from −0.926 to 1.083. Higher value means more critical to give an affirmative response to the desirable brand equity statement. The threshold estimates of items and dimensions are shown in Tables 3 and 4. An item/dimension receiving a high raw score suggests that this item/dimension has higher threshold for consumers to give affirmative response. The range of dimension estimates is from −0.485 for brand awareness to 0.119 for uniqueness. That is, brand awareness is the easiest dimension to agree with whereas uniqueness is the most difficult one. Item

**Table 1:** Brand equity measures from MFIRT (master brands and sub-brands)

| *Master brand* | *Master brand equity estimate* | *Sub-brand* | *Sub-brand individual equity estimate* | *Sub-brand equity* |
|---|---|---|---|---|
| Coca-Cola | 1.000 | Classic Coke | 0.266 | 1.266 |
| | | Coke Zero | −0.079 | 0.921 |
| | | Diet Coke | −0.187 | 0.813 |
| Pepsi | 0.807 | Original Pepsi | 0.229 | 1.036 |
| | | Diet Pepsi | −0.055 | 0.752 |
| | | Diet Pepsi Max | −0.174 | 0.634 |

*Note*: Sub-brand equity is calculated as the sum of master brand equity estimate and sub-brand individual equity estimate.

**Table 2:** MFIRT analysis of consumers – Brand portfolio

| *Consumer* | *Criticality estimate* | *SD* | *2.5%* | *97.5%* |
|---|---|---|---|---|
| #1 | 1.083 | 0.054 | 0.976 | 1.186 |
| #2 | 0.964 | 0.051 | 0.860 | 1.063 |
| #3 | 0.730 | 0.049 | 0.639 | 0.826 |
| #4 | 0.662 | 0.049 | 0.565 | 0.760 |
| #5 | 0.629 | 0.049 | 0.532 | 0.725 |
| #6 | −0.606 | 0.052 | −0.708 | −0.501 |
| #7 | −0.615 | 0.053 | −0.723 | −0.510 |
| #8 | −0.622 | 0.052 | −0.729 | −0.521 |
| #9 | −0.831 | 0.054 | −0.942 | −0.729 |
| #10 | −0.926 | 0.057 | −1.039 | −0.813 |

*Note*: 2.5 and 97.5 per cent are percentile scores, which can be interpreted as confidence interval.

The estimates of 10 randomly selected respondents are reported here. Details about other respondents can be obtained from the authors.

**Table 3:** MFIRT analysis of items – Brand portfolio

| *Dimension* | *Item* | *Estimate* | *SD* | *2.5%* | *97.5%* |
|---|---|---|---|---|---|
| Brand awareness | I have heard of this brand. | −0.557 | 0.021 | −0.599 | −0.515 |
| | I am aware of this brand. | 0.274 | 0.017 | 0.240 | 0.307 |
| | I am very familiar with this brand. | −0.318 | 0.020 | −0.357 | −0.277 |
| | I have an opinion about this brand. | 0.596 | 0.016 | 0.563 | 0.628 |
| | When I think of the product category this brand belongs to, this brand is the first comes to mind. | 0.004 | 0.018 | −0.030 | 0.042 |
| Perceived quality | The likelihood that this brand is reliable is very high. | −0.048 | 0.017 | −0.081 | −0.016 |
| | The quality of this brand is very high. | 0.011 | 0.016 | −0.021 | 0.042 |
| | In terms of overall quality, I'd rate this brand high. | 0.056 | 0.016 | 0.025 | 0.088 |
| | I can always count on this brand for consistent high quality. | 0.066 | 0.016 | 0.034 | 0.097 |
| | This brand is a quality leader within its product category. | −0.085 | 0.016 | −0.118 | −0.053 |
| Brand loyalty | I would buy the brand on the next opportunity. | −0.097 | 0.016 | −0.128 | −0.065 |
| | I would recommend the product or service to others. | 0.014 | 0.016 | −0.019 | 0.044 |
| | X would be my first choice. | −0.118 | 0.016 | −0.148 | −0.085 |
| | I will not buy other brands if X is available at the store. | 0.019 | 0.016 | −0.011 | 0.049 |
| | The next time, I buy (product category), I intend to buy a (brand name) brand. | 0.182 | 0.016 | 0.151 | 0.211 |
| Brand associations | I can quickly recall the symbol or logo of this brand. | 0.174 | 0.016 | 0.142 | 0.206 |
| | This brand is a very good brand | −0.421 | 0.019 | −0.457 | −0.385 |
| | This brand is a very nice brand. | 0.047 | 0.016 | 0.015 | 0.079 |
| | This brand is an extremely likeable brand. | 0.071 | 0.016 | 0.040 | 0.105 |
| | I have a clear image of the type of person who would use the brand. | 0.130 | 0.016 | 0.098 | 0.161 |
| Uniqueness | This brand is 'distinct' from other brands in the same product category. | 0.037 | 0.015 | 0.007 | 0.067 |
| | This brand is very different from other brands in the same product category. | −0.024 | 0.016 | −0.055 | 0.008 |
| | This brand really stands out from other brands in the same product category. | −0.044 | 0.015 | −0.073 | −0.014 |
| | This brand is 'unique' from other brands in the same product category. | 0.008 | 0.015 | −0.024 | 0.038 |
| | This brand is different from competing brands. | 0.023 | 0.016 | −0.006 | 0.054 |

*Note*: 2.5 and 97.5 per cent are percentile scores, which can be interpreted as confidence interval.

**Table 4:** MFIRT analysis of dimensions – Brand portfolio

| *Dimension* | *Estimates* | *SD* | *2.5%* | *97.5%* |
|---|---|---|---|---|
| Brand awareness | −0.485 | 0.009 | −0.502 | −0.468 |
| Perceived quality | −0.209 | 0.008 | −0.223 | −0.194 |
| Brand loyalty | 0.010 | 0.008 | −0.004 | 0.025 |
| Brand associations | −0.265 | 0.008 | −0.281 | −0.250 |
| Uniqueness | 0.119 | 0.007 | 0.104 | 0.133 |

*Note*: 2.5 and 97.5 per cent are percentile scores, which can be interpreted as confidence interval.

estimates also vary within and across dimensions. The step calibrations are in ascending order (that is, −0.51, −0.33, −0.15, −0.06, 0.12, 0.33 and 0.61) implying that as one brand moves up the brand equity continuum (or one consumer moves up to be less critical), each category in turn becomes the most probable response.[7] Thus, the Likert scale functions well in the current study.

## DISCUSSION AND IMPLICATION

Followers of the traditional measurement paradigm[23] implicitly assume CBBE scales have only one purpose – scaling people. While common in psychology, it is not always appropriate in marketing especially for CBBE measurement. Wang and Finn[6] introduced MGT and developed a brand level scale to fill this gap. It is designed to maximally differentiate brands from each other and reliably capture their differences. This is the first scale that has been developed for CBBE measurement area at the brand level. MGT approach introduced in their study quantifies all the design factors involved and increases the validity and reproducibility of the CBBE estimates. However, MGT has little interest in the brand equity reporting *per se*. Aggregate evaluations (such as means) are still used for CBBE comparison among brands. An average is a joint property of all dimensions and items in a particular scale and the particular individuals sampled. When dimensions (and items) are added to or dropped from a scale, or respondents change, the average will change. Therefore, the average CBBE at the brand level could be at best contextually biased and at worst an unreliable and invalid estimate. The average also inappropriately assumes the responses are interval-scaled instead of ordinal-scaled as they really are. MFIRT analysis addresses the limitations of this aggregate CBBE evaluation approach. It models brand equity (at the brand level) as a separate term, independent of particular dimensions, items and individuals involved in the evaluation. The ordinal nature of the rating scale is also examined. This model characteristic of independence not only provides better estimates of brand equity but also offers flexibility in terms of data collection.

The present research has several implications. First, it provides a complementary framework for researchers to further understand and measure CBBE beyond the CTT and MGT approaches. It is a new and comprehensive perspective and offers a lot of opportunities to be further developed (for example, examining antecedents and consequences at either consumer or brand level) and adapted to other research questions (such as scaling the interactions between brands and consumers). Second, the proposed scale increases the objectivity and managerial relevance of CBBE. Managers can use this newly developed scale to get reliable brand estimates of CBBE that can be compared with the results from other brand equity measurement approaches (that is, finance and economics approaches) to identify the existence of inconsistency and investigate the reason. It helps managers precisely monitor the response from the consumer side on their brand marketing moves, the crucial step leading to the success of any marketing program, and facilitates timely adjustments. For managers who are interested in the relationships between CBBE and other marketing strategic variables, this work provides unbiased estimates that have high reliability and validity when scaling brands. Using these high-quality brand CBBE estimates as inputs, the correlations they get and the inference they make about the relationships are unbiased. Third, by explicitly separating brand parameters from those for consumers, dimensions and items, this research makes effort to bridge the gap between objective measures and subjective measures in the brand equity measurement area. Researchers are interested in the 'true' scores of variables of interest (CBBE). However, they only have observed scores, which are scores that confound the true scores with multiple extraneous sources of variation such as the different response patterns of individual consumers. Some of them may have either lower or higher probability to use extreme response categories (such as 1s and 7s in a 1-7 Likert scale). The current research not only conceptually separates the true value from other sources of variance but also introduces the methodologies into marketing (that is, MFIRT) that enables the separation. Finally, the current research provides a tool for brand portfolio managers to measure the CBBE of their master brands and sub-brands and also offers useful information about the comparative strength of the master brands and their sub-brands.

This research has some limitations that we plan to address in future research. The nature of the dimensions (that is, formative versus reflective) and the possible interactions (such as interactions between brands and consumers) are not recognized in the current MFIRT analysis. The nature of the dimensions can be examined by testing a MFIRT measurement model that incorporates a hierarchical structure. Substantial interactions between brands and consumers are commonly observed in marketing literature and practice. An interaction term can be introduced into a CBBE MFIRT measurement model to quantify this effect.

## REFERENCES


1 Bharadwaj, S.G., Varadarajan, P.R. and Fahy, J. (1993) Sustainable competitive advantage in service industries: A conceptual model and research propositions. *Journal of Marketing* 57(4): 83–99.
2 Yoo, B. and Donthu, N. (2001) Developing and validating a multidimensional consumer-based brand equity scale. *Journal of Business Research* 52(1): 1–14.
3 Netemeyer, R.G. *et al* (2004) Developing and validating measures of facets of customer-based brand equity. *Journal of Business Research* 57(2): 209–224.
4 Aaker, D.A. (1991) *Managing Brand Equity: Capitalizing on the Value of a Brand Name*. New York: The Free Press.
5 Keller, K.L. (1993) Conceptualizing, measuring, and managing customer-based brand equity. *Journal of Marketing* 57(1): 1–22.
6 Wang, L. and Finn, A. (2012) Measuring consumer-based brand equity across brand portfolios: Generalizability theory perspective. *Journal of Targeting, Measurement and Analysis for Marketing* 20: 109–116.
7 Smith Jr, E.V. and Kulikowich, J.M. (2004) An application of generalizability theory and many-facet Rasch measurement using a complex problem-solving skills assessment. *Educational and Psychological Measurement* 64(4): 617–639.
8 Aaker, D.A. (2004) *Brand Portfolio Strategy: Creating Relevance, Differentiation, Energy, Leverage, and Clarity*. New York: Simon and Schuster.
9 Bradlow, E.T. and Thomas, N. (1998) Item response theory models applied to data allowing examinee choice. *Journal of Educational and Behavioral Statistics* 23(3): 236–243.
10 Embretson, S.E. (2002) Generating abstract reasoning items with cognitive theory. In: S.H. Irvine and P. Kyllonen (eds.) *Item Generation for Test Development*. Mahwah, NJ: Lawrence Earlbaum Associates, pp. 219–250.
11 Balasubramaniam, S.K. and Kamakura, W.A. (1989) Measuring consumer attitudes toward the marketplace with tailored interviews. *Journal of Marketing Research* 26(3): 311–326.
12 Singh, J., Howell, R.D. and Rhoads, G.K. (1990) Adaptive designs for Likert-type data: An approach for implementing marketing surveys. *Journal of Marketing Research* 27(3): 304–321.
13 de Jong, M.G., Steenkamp, J.B.E.M. and Fox, J.P. (2007) Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *Journal of Consumer Research* 34(2): 260–278.
14 de Jong, M.G., Steenkamp, J.B.E.M., Fox, J.P. and Baumgartner, H. (2008) Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research* 45(1): 104–115.
15 de Jong, M.G., Steenkamp, J.B.E.M. and Veldkamp, B.P. (2009) A model for the construction of country-specific yet internationally comparable short-form marketing scales. *Marketing Science* 28(4): 674–689.
16 Linacre, J.M. (1989) *A User's Guide to Facets: Rasch Measurement Computer Program*. Chicago, IL: Mesa.
17 Andrich, D. (1978) A rating formulation for ordered response categories. *Psychometrika* 43(4): 561–573.
18 Rossi, P.E., Gilula, Z. and Allenby, G.M. (2001) Overcoming scale usage heterogeneity: A Bayesian hierarchical approach. *Journal of the American Statistical Association* 96(453): 20–31.
19 Van der Linden, W.J. and Hambleton, R.K. (1997) Item response theory: Brief history, common models, and extensions. In: W.J. van der Linden and R.K. Hambleton (eds.) *Handbook of Modern Item Response Theory*. New York: Springer-Verlag, pp. 1–28.
20 Lumley, T. and McNamara, T.F. (1995) Rater characteristics and rater bias: Implications for training. *Language Testing* 12(1): 54–71.
21 Gelman, A. (2006) Prior distribution for variance parameters in hierarchical models (Comment on Article by Browne and Draper). *Bayesian Analysis* 1(3): 515–534.
22 Brooks, S.P. and Gelman, A. (1998) General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* 7(4): 434–455.
23 Churchill, G.A. (1979) A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research* 16(1): 64–73.