Cross-Cultural Analysis of Gallup's Q12® Employee Engagement Instrument

James K. Harter, Ph.D., and Sangeeta Agrawal, M.S.

GALLUP CONSULTING®

For more information about Gallup Consulting or our solutions for optimizing business performance, please visit consulting.gallup.com or contact Sarah Van Allen at 202.715.3152 or sarah_van_allen@gallup.com.

Cross-Cultural Analysis of Gallup's Q12® Employee Engagement Instrument

By James K. Harter, Ph.D., and Sangeeta Agrawal, M.S.

Executive Summary

Objective

To review past research and present new research examining the cross-cultural application of Gallup's Q12 employee engagement instrument

Methods

Gallup researchers examined a variety of statistical indices across more than 170,000 work units and 28 language-country combinations to assess any possible differential functioning of the instrument across international settings. They reviewed factor analysis, reliability, validity, and differential item functioning (DIF) indices.

Results

The various analyses indicated that the instrument functions similarly across the different contexts studied and across the many different translations. Researchers observed mean differences across some language-country combinations, but found no evidence of bias (non-uniform DIF) across 28 language-country combinations. These results are consistent with prior research on the instrument.

Conclusion

All evidence suggests the Q12 employee engagement instrument has properties that support its generalizable use across a variety of worldwide contexts. Given that researchers noted some mean differences for certain items and language-country combinations, the most conservative use of Gallup's database is to compare one's work unit engagement data with local country-language data from the database. Translation is not an exact science, and it is important that future research continue to monitor and consider improvements to translations based on the best possible expert judgment, along with statistical guidance.

Gallup's Q12 employee engagement instrument has had significant historical development and cross-cultural testing. Gallup designed the instrument to measure generalizable workplace elements that can be managed to improve the performance outcomes of organizations. Because the instrument is a standardized measure that Gallup applies in a variety of settings, it is important that it is studied and evaluated across those settings. Gallup's 12 core employee engagement items are as follows:

Q01. I know what is expected of me at work.

Q02. I have the materials and equipment I need to do my work right.

Q03. At work, I have the opportunity to do what I do best every day.

Q04. In the last seven days, I have received recognition or praise for doing good work.

Q05. My supervisor, or someone at work, seems to care about me as a person.

Q06. There is someone at work who encourages my development.

Q07. At work, my opinions seem to count.

Q08. The mission or purpose of my organization makes me feel my job is important.

Q09. My associates or fellow employees are committed to doing quality work.

Q10. I have a best friend at work.

Q11. In the last six months, someone at work has talked to me about my progress.

Q12. This last year, I have had opportunities at work to learn and grow.

Summary of Past Research

Gallup has conducted several studies examining issues of cross-cultural applications of the Q^{12} instrument, as administered across different countries and languages. These analyses include overall scale functioning, item functioning, and criterion-related validity studies. Decades of early research included application and examination of the instrument qualitatively and quantitatively across a variety of countries (development in Harter, Schmidt, Killham, & Agrawal, 2009). Additional in-depth research has been conducted more recently:

Cross-Cultural Research on Gallup's Q^{12} (2001). Gallup research paper. Omaha, NE.

Based on data collected from Gallup's client database, researchers studied the item characteristic curves across different regions of the world. Findings indicated that, while percentage positive response to each item varies across different parts of the world, each item functions (relates to overall engagement) in a similar manner across regions.

Employee Engagement Cross-Cultural Research (2003). Gallup research paper. Omaha, NE.

Gallup collected random samples of working populations from a series of workplace polls in the U.S., U.K., Singapore, New Zealand, Australia, Japan, Israel, Germany, and France. Gallup's engagement index was studied in relationship to intent to stay and likelihood to recommend the organization to others. While the percent engaged varied by country, the relationship between engagement and these outcomes remained consistent across the countries studied. Via meta-analysis, the relationship among overall engagement (GrandMean of Q^{12}), the engagement index (engaged, not engaged, and actively disengaged categories), and the outcome variables was essentially the same across countries.

Sireci, S., Harter, J., Yang, Y., & Bhola, D. (2003). Evaluating the Equivalence of an Employee Attitude Survey Across Languages, Cultures, and Administration Formats. International Journal of Testing, 3(2), 129-150.

The study used weighted multidimensional scaling (MDS), analysis of covariance (ANCOVA), and ordinal logistic regression (LR) to evaluate the structural equivalence and DIF of the Q^{12} and other survey items. Researchers evaluated items across three different languages and eight cultures. Results indicated the structure of the survey data was consistent and that items functioned similarly across all groups.

Sireci, S., Yang, Y., Harter, J., & Ehrlich, E. (2006). Evaluating Guidelines for Test Adaptions: A methodological analysis of translation quality. Journal of Cross-Cultural Psychology, 37(5), 557-567.

Researchers summarized standards for increasing translation quality and used DIF methodology to evaluate the comparability of translated items at two different points in time (after the original and revised translations). Statistical results indicated improvements in many items due to translation revisions. Overall, the methods illustrated how different translations can be evaluated statistically, in addition to conceptual judgment.

Harter, J., Schmidt, F., Killham, E., & Agrawal, S. (2009). Q^{12} Meta-Analysis: The relationship between engagement at work and organizational outcomes. Gallup Technical Report. Omaha, NE.

Researchers accumulated 199 studies across 152 organizations in 44 industries and 26 countries to assess the relationship between the Q^{12} employee engagement instrument and various performance outcomes, including customer loyalty/engagement, profitability, productivity, turnover, safety incidents, shrinkage, absenteeism, and quality (defects). The meta-analysis included research studies from Asia (Hong Kong, Japan, Korea, and Thailand), Australia, New Zealand, Europe (Netherlands, Germany, United Kingdom, Ireland, France, Austria, Italy, Spain, Belgium, and Greece), former communist countries (Russia, Hungary, Lithuania, Czech Republic, and Poland),

Latin America (Brazil and Mexico), the Middle East (United Arab Emirates), and North America (Canada and the United States). Findings of the meta-analysis and validity generalization research indicated the relationship between the Q^{12} and various performance outcomes was meaningful and highly generalizable across the different situations studied.

Harter, J., & Agrawal, S. (2010). A worldwide study of employee engagement and its relationship to wellbeing, health, and giving. Gallup Technical Report. Omaha, NE.

Based on data collected across 119 countries from the Gallup World Poll, researchers studied the relationship between the Q^{12} instrument and wellbeing, health, and giving. Findings of the meta-analysis and validity generalization research indicated that the general direction of the relationship between engagement and wellbeing, health, and giving was consistent across countries.

Objective of the Present Study

The purpose of this research is to provide an updated analysis that examines possible differential functioning of the Q^{12} across language-country combinations using multiple methods, including:

- Factor analysis to assess consistency of the factor structure
- Reliability analysis and correlation of each item to the overall engagement scale to assess the internal consistency of the instrument
- DIF analysis, using hierarchical multiple regression (HMR) and ANCOVA methods

Any differences found can be used to inform use of the Q^{12} instrument in different international situations and in directing resources for systematic translation review.

Methods

Translations

Gallup has historically followed a detailed 18-step translation process that can be condensed into the following five general steps:

1. Finalize the English/U.S. item and amplified English source (representing the psychology of the item).
2. Translate into target language; perform independent back-translation.
3. Conduct multiple reviews of English/U.S. item and amplified English source.
4. Finalize target language version.
5. Local expert performs final test and provides feedback on language and content.

Analyses

The present set of analyses uses work unit-level data from Gallup's employee engagement client database to assess scale properties. Gallup reports Q^{12} scores at the workgroup level within organizations, so it is appropriate to study them at the level in which they are reported and interpreted. The overall database, dating back to 1996, now includes approximately 20 million respondents and 2.2 million work units in 68 languages and 183 countries. To maximize information, Gallup chose to study country-language combinations and used various inclusion criteria, as follows:

1. Because many organizations within the overall database include multiple administrations of Q^{12}, and Q^{12} scores are higher for second administrations and beyond, researchers initially selected 436,179 first administration work units.
2. To ensure homogeneity of language and country within workgroups, researchers further reduced the sample to 408,708 work units where 99% or more

of the work unit respondents were from the same country and the survey was administered in the same language.

3. As work units vary by size, and size of work unit can influence variability, researchers selected 208,174 work units containing from 5 to 50 respondents.

4. To use relatively recent data, researchers chose 198,710 work units from 2000-2010.

5. Researchers selected a minimum sample size of 300 work units per country-language combination to maximize the stability of outcomes, resulting in 28 language-country combinations and reducing the overall sample size to 170,508 work units.

The above criteria resulted in 16 languages and 23 countries, or 28 language-country combinations. Sample sizes varied from 115,166 work units responding in English/U.S. to 324 for English/Hong Kong.

Gallup examined differential functioning using multiple methods, including:

Factor analysis: Prior research has found the Q^{12} to contain a dominant first factor explaining most of the variance in the measure. One assessment of differential functioning involved calculating the percentage of variance accounted for by the first factor using principal components factor analysis.

Reliability analysis: The Cronbach's alpha reliability, in addition to test-retest reliability, has been extensively reported in prior publications. Another test of differential functioning is to compare the Cronbach's alpha reliabilities across the 28 language-country combinations.

Correlation of each item with the overall engagement scale: A third test involved examining the consistency of relationship between each item and the overall engagement scale (each item correlated to the total minus the focal item, correcting for part-whole overlap) across the 28 language-country combinations.

DIF using HMR and ANCOVA methods: DIF analysis is aimed toward identifying items that function differently across various language versions. Several statistical methods for evaluating DIF are available (Sireci & Berberoglu, 2000; Thissen, Steinberg, & Wainer, 1993; Zumbo, 1999). Most DIF analysis methods are essentially variants on traditional ANCOVA methods, which evaluate group differences on an outcome measure after controlling for variables that may distort interpretation of the differences.

Controlling for management status (executive, manager, and non-manager), job function (nine major functions), and industry type, researchers employed two methods of estimating DIF. They used the English/U.S. version of the questionnaire (the initial version) as the base language-country combination to which each of the remaining 27 language-country combinations was compared, resulting in 324 analyses (12 items x 27 language-country comparisons). Because the sample size was proportionately much greater for the English/U.S. version of the instrument, researcher drew a random sample of English/U.S. work units (from the overall sample of 115,166 work units) equal to the number of work units in each language/country-paired comparison.

HMR and ANCOVA methods yielded identical results, so this paper reports only the HMR analyses. In the analyses, researchers treated each Q^{12} item as a dependent variable. Independent variables were entered as steps:

1. The primary covariate, GrandMean minus the focal item, centered to eliminate multicollinearity concerns, in addition to manager status, job function, and industry type (dummy coded)

2. The language country group (English/U.S.=1, focal country-language=0)

3. The covariate x group interaction term

The R squared change statistic is used to assess DIF along a three-category continuum (small, medium, or large DIF).

DIF Cutoffs: R-Square Change	
small	< 0.035
medium	.035-.069
large	.07+

These analyses resulted in two primary statistics that can be categorized as indicating small, medium, or large DIF. The first index assesses non-uniform DIF, which indicates whether there are differences in the relationship between the item and engagement across different languages (across different engagement levels). This is determined by calculating the R-square change statistic for step 3 minus step 2.

The second index assesses uniform DIF, which indicates if the mean scores for the different languages are uniformly different across engagement levels. This is determined by calculating the R-square change statistic for step 2 minus step 1.

Non-uniform DIF is most important in indicating bias. Uniform DIF may indicate mean scores are different, but consistently different across engagement levels, which may reflect either the true difference in respondents' engagement levels or a difference in the item's difficulty based on its translation. Non-uniform DIF indicates the most severe problem, and uniform DIF should be interpreted in the context of other information about the translations.

Figures 1-3 (see page 10) provide visual examples of non-uniform DIF, uniform DIF, and no DIF.

Results

Table 1 (see page 7) provides the total number of work units, respondents, and respondents per work unit, descriptive statistics, data collection time periods, factor analysis, and reliability statistics. Researchers observed the highest scores on the composite engagement measure in Korean/Korea and Indonesian/Indonesia languages. They observed the lowest scores in Italian/Italy, French/France, and Japanese/Japan. It is important to note that the data do not contain random samples of work units within the language/country combinations. As such, any differences in mean scores could be due to the types of work units sampled within each language/country combination.

The percentage of variance explained by the first factor approached or exceeded 50% in all language/country combinations (mean of 56%, with range of 46%-79%), and reliabilities were very high in all cases (mean of .92).

Table 2 (see page 8) provides the correlation of each item to the overall Q^{12} GrandMean. Each item correlated positively with the overall scale score (corrected for part-whole overlap). This is a positive indication that each Q^{12} item is consistent with the other items in the scale.

Researchers conducted DIF analysis using the HMR procedure. This involved running 324 regression analyses. None contained evidence of medium or large non-uniform DIF, meaning no systematic bias was discovered across the translation versions. Researchers found 13 instances (4%) of medium uniform DIF and 10 instances (3%) of large uniform DIF. In other words, the mean scores were systematically different across the engagement continuum in these rare cases, but the lines didn't cross. To picture this visually, see Figures 4 and 5 (pages 10 and 11). Figure 4 is an example of uniform DIF on item 10 (I have a best friend at work), and Figure 5 is an example of no uniform DIF on the same item. Large uniform DIF was most likely to occur on the translation of "I have a best friend at work." The non-English version of this item tended to produce higher mean scores than the English version.

	Total	Large	Medium	Small
Non-Uniform DIF	324	0	0	324
Uniform DIF	324	10	13	301

Overall, results of the DIF analysis suggest no systematic bias in the functioning of the Q^{12} items across translations; however, there is some evidence of uniform DIF. In addition to Q10, researchers detected uniform DIF in a small number of language/country combinations for Q01 (I know what is expected of me at work), Q05 (My

supervisor, or someone at work, seems to care about me as a person), Q11 (In the past six months, someone at work has talked to me about my progress), and Q03 (At work, I have the opportunity to do what I do best every day). It is important to note that there have been iterations of translation improvements in some items for some languages. The present analysis pools together all available data. A next step in this process will be to separate different translation iterations and study the relative difference in DIF statistics among various iterations.

Conclusion

Upon review of this present research and all prior cross-cultural research on the Q^{12} employee engagement instrument, strong evidence continues to show that it is appropriate to use the instrument across a wide variety of international settings. The instrument was historically developed and has been tested and applied in varied international settings. Researchers have conducted follow-up studies using a variety of methods, including standard psychometric analysis (factor analysis, reliability analysis, and part-whole correlations), ANCOVA, and HMR. These efforts have demonstrated that the instrument, as a whole, functions in a similar manner across different language-country combinations. Overall, language translations are effective and robust. Translation combines qualitative knowledge of the local language and culture, along with quantitative data, but it is not a perfect science. Some translations undergo refinements based on changes in local language use and feedback from local experts. This is an ongoing process, and it is expected that continued improvements can be made in some situations. However, this present analysis, along with past analysis, strongly indicates the current instrument does not produce biased results across languages and countries. On this point, the evidence is strong. Past research has also shown that the instrument consistently predicts important outcomes across different settings, suggesting the scores produced from the Q^{12} have wide applicability around the world.

References

Sireci, S., & Berberoglu, G. (2000). Using bilingual respondents to evaluate translated-adapted items. *Applied Measurement in Education, 13*(3), 229-248.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation. Department of National Defense.

Table 1

Descriptive Statistics, Factor and Reliability Analysis

Language/Country	Total Number of Work Units	Total Number of Respondents	Number of Respondents/ Work Unit (Range 5-50) Mean	GrandMean Work Unit Level Z*	Year Range	Factor Analysis Variability Accounted by First Factor	Reliability Cronbach's Alpha (12 Items)
0) English/U.S.	115,166	1,296,985	11	.09	2000-2010	56.47	.925
1) English/U.K.	8,615	96,103	11	-.40	2000-2010	56.66	.927
2) English/India	11,858	121,972	10	.00	2000-2010	58.88	.934
3) English/Canada	4,732	58,707	12	-.22	2000-2010	57.02	.925
4) German/Germany	5,656	71,715	13	-.42	2000-2010	52.24	.911
5) Chinese Simplified (Mandarin)/China	2,791	26,734	10	.11	2001-2010	59.64	.937
6) English/Indonesia	4,148	40,798	10	-.11	2000-2010	48.59	.896
7) English/Australia	2,865	27,170	9	-.44	2000-2010	52.79	.912
8) Thai/Thailand	3,242	29,452	9	.20	2004-2010	78.65	.974
9) English/Germany	1,265	11,420	9	-.22	2003	52.18	.909
10) English/Singapore	789	8,320	11	-.22	2000-2010	56.85	.929
11) Japanese/Japan (including Okinawa)	587	6,195	11	-.58	2001-2010	55.26	.924
12) French (Euro)/France	606	6,243	10	-.64	2000-2010	53.87	.906
13) Polish/Poland	727	6,630	9	-.09	2006-2010	62.75	.942
14) Italian/Italy	839	8,737	10	-.71	2004-2010	55.5	.922
15) Portuguese (Intl.)/Brazil	929	11,129	12	.27	2004-2010	62.61	.942
16) Indonesian/Singapore	307	2,258	7	.11	2008	59.43	.935
17) Spanish (Latin Am)/U.S.	502	5,733	11	-.24	2002-2008	58.76	.930
18) Cantonese/China	470	4,013	9	-.49	2002	51.67	.911
19) Korean/Korea	649	6,585	10	.47	2002-2010	61.04	.941
20) Spanish (Latin Am)/Mexico	343	3,253	9	.09	2006-2010	52.03	.911
21) English/South Africa	326	2,540	8	.07	2005-2010	45.9	.884
22) English/Hong Kong	324	3,929	12	-.42	2000	53.62	.919
23) French (Canadian)/Canada	334	3,908	12	-.09	2007	53.39	.909
24) Dutch/Netherlands	440	3,751	9	-.42	2004-2010	46.45	.882
25) Swedish/Sweden	352	3,541	10	-.33	2008-2010	51.86	.905
26) Indonesian/Indonesia	1,274	14,610	11	.44	2004-2010	64.41	.948
27) English/United Arab Emirates	372	3,589	10	-.11	2001-2010	51.7	.910
TOTAL	170,508	1,886,020	11		2000-2010	56.08	.922

*Work unit-level standard deviation

Table 2

Correlation of Each Item With Q12® GrandMean*												
Language/Country	**Q01**	**Q02**	**Q03**	**Q04**	**Q05**	**Q06**	**Q07**	**Q08**	**Q09**	**Q10**	**Q11**	**Q12**
0) English/U.S.	.591	.612	.714	.745	.790	.845	.813	.733	.577	.339	.673	.805
1) English/U.K.	.598	.597	.674	.758	.802	.865	.807	.721	.478	.432	.724	.827
2) English/India	.677	.581	.799	.742	.752	.831	.743	.765	.663	.512	.727	.764
3) English/Canada	.574	.611	.714	.754	.807	.870	.827	.767	.554	.236	.728	.791
4) German/Germany	.387	.533	.578	.737	.776	.851	.792	.681	.611	.362	.706	.771
5) Chinese Simplified (Mandarin)/ China	.662	.588	.752	.767	.786	.833	.816	.711	.722	.471	.776	.739
6) English/Indonesia	.483	.538	.595	.680	.753	.812	.757	.652	.431	.305	.629	.739
7) English/Australia	.559	.575	.703	.726	.754	.820	.768	.734	.507	.309	.622	.766
8) Thai/Thailand	.848	.802	.890	.907	.878	.906	.909	.878	.828	.753	.875	.889
9) English/Germany	.546	.546	.640	.690	.744	.828	.770	.749	.594	.354	.606	.726
10) English/Singapore	.660	.639	.763	.752	.731	.817	.716	.704	.663	.513	.672	.731
11) Japanese/Japan (including Okinawa)	.727	.521	.696	.730	.706	.747	.772	.745	.631	.492	.664	.754
12) French (Euro)/France	.616	.487	.623	.752	.764	.827	.773	.775	.519	.145	.743	.774
13) Polish/Poland	.600	.657	.749	.797	.838	.872	.827	.818	.769	.494	.759	.727
14) Italian/Italy	.618	.549	.701	.749	.742	.857	.811	.744	.544	.277	.755	.768
15) Portuguese (Intl.)/Brazil	.711	.639	.823	.800	.815	.873	.778	.778	.646	.468	.790	.775
16) Indonesian/Singapore	.652	.646	.804	.672	.704	.760	.821	.759	.727	.455	.779	.814
17) Spanish (Latin Am)/U.S.	.557	.630	.761	.811	.803	.872	.819	.819	.491	.320	.740	.803
18) Cantonese/China	.611	.500	.618	.670	.748	.796	.744	.672	.702	.413	.606	.705
19) Korean/Korea	.657	.644	.742	.784	.792	.827	.807	.735	.684	.615	.751	.751
20) Spanish Latin Am)/Mexico	.529	.457	.653	.727	.759	.832	.764	.681	.519	.356	.733	.765
21) English/South Africa	.434	.493	.637	.716	.697	.751	.719	.639	.480	.231	.559	.690
22) English/Hong Kong	.612	.411	.655	.704	.755	.774	.719	.749	.626	.537	.712	.748
23) French (Canadian)/Canada	.563	.518	.547	.722	.763	.847	.806	.762	.566	.293	.768	.709
24) Dutch/Netherlands	.477	.475	.605	.702	.732	.742	.747	.649	.484	.112	.559	.737
25) Swedish/Sweden	.553	.531	.630	.743	.768	.859	.767	.708	.414	.239	.684	.773
26) Indonesian/Indonesia	.671	.721	.808	.753	.796	.815	.822	.780	.766	.626	.773	.776
27) English/United Arab Emirates	.571	.525	.741	.696	.713	.778	.694	.702	.578	.390	.674	.721

*Each item was correlated with the GrandMean minus the focal item to correct for part-whole overlap

Table 3

HMR Method	Uniform DIF	
Compared With English/U.S.	Large	Medium
1) English/U.K.		
2) English/India	Q10	
3) English/Canada		
4) German/Germany		Q01
5) Chinese Simplified (Mandarin)/China	Q10	Q05
6) English/Indonesia		
7) English/Australia		
8) Thai/Thailand	Q10	Q05, Q11
9) English/Germany		
10) English/Singapore	Q10	
11) Japanese/Japan (including Okinawa)		Q05
12) French (Euro)/France		
13) Polish/Poland		
14) Italian/Italy		Q10
15) Portuguese (Intl.)/Brazil		Q10
16) Indonesian/Singapore		Q10
17) Spanish (Latin Am)/U.S.		
18) Cantonese/China		
19) Korean/Korea	Q10	
20) Spanish (Latin Am)/Mexico	Q03, Q10	Q05, Q11
21) English/South Africa		
22) English/Hong Kong	Q01, Q10	
23) French (Canadian)/Canada		
24) Dutch/Netherlands	Q10	
25) Swedish/Sweden		
26) Indonesian/Indonesia		Q10, Q11
27) English/United Arab Emirates		Q10

Figure 1

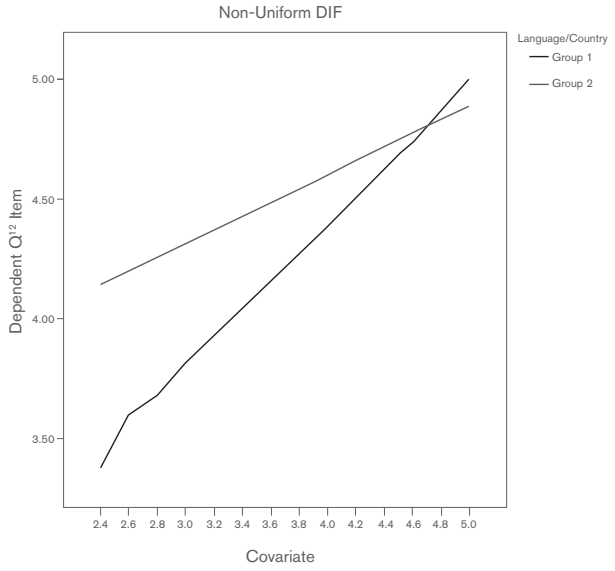


Figure 2

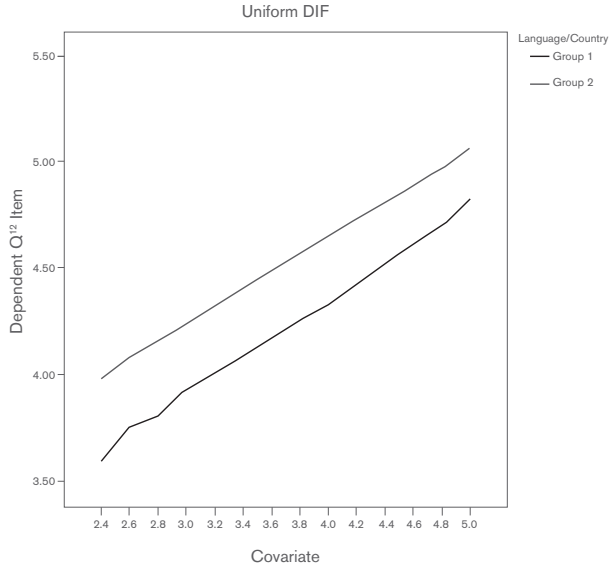


Figure 3

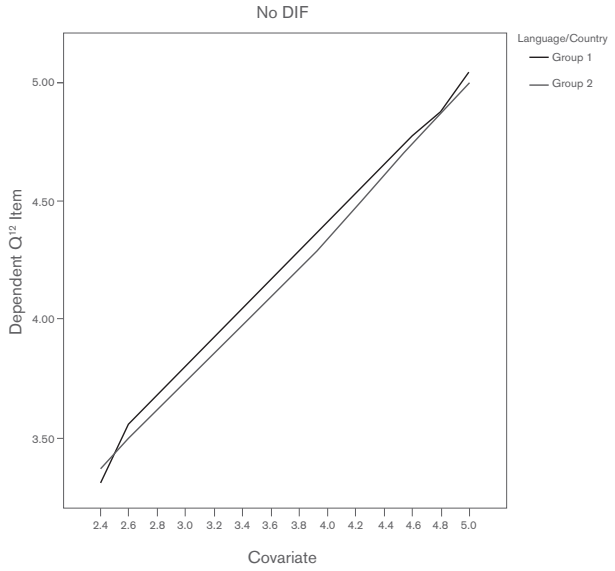


Figure 4

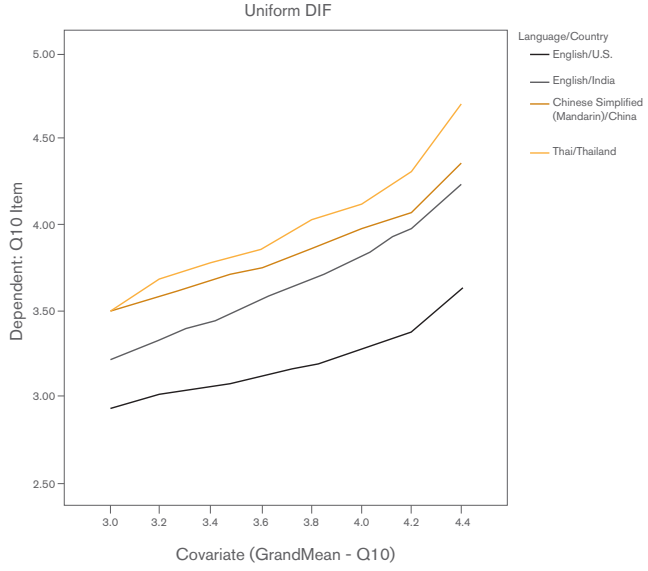
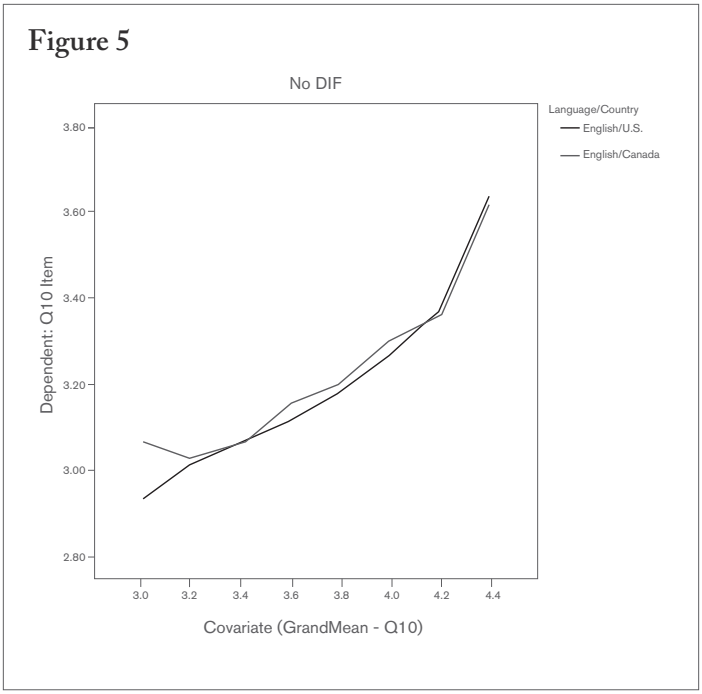


Figure 5



GALLUP®

World Headquarters

The Gallup Building
901 F Street, NW
Washington, D.C. 20004

t +1.877.242.5587
f +1.202.715.3045

www.gallup.com