3-2009

Investigating the Comparability of a Self-Report Measure of Childhood Bullying Across Countries

Chiaki Konishi
University of British Columbia

Shelley Hymel
University of British Columbia

Bruno D. Zumbo
University of British Columbia

Zhen Li
University of British Columbia

Mitsuru Taki
National Institute for Educational Policy Research, Tokyo

See next page for additional authors

Konishi, Chiaki; Hymel, Shelley; Zumbo, Bruno D.; Li, Zhen; Taki, Mitsuru; Slee, Phillip; Pepler, Debra; Sim, Hee-og; Craig, Wendy; Swearer Napolitano, Susan M.; and Kwak, Keumjoo, "Investigating the Comparability of a Self-Report Measure of Childhood Bullying Across Countries" (2009). *Educational Psychology Papers and Publications*. 151.
https://digitalcommons.unl.edu/edpsychpapers/151

Authors

Chiaki Konishi, Shelley Hymel, Bruno D. Zumbo, Zhen Li, Mitsuru Taki, Phillip Slee, Debra Pepler, Hee-og Sim, Wendy Craig, Susan M. Swearer Napolitano, and Keumjoo Kwak

Published in the *Canadian Journal of School Psychology* (March 2009) 24(1): 82-93. Copyright 2009, SAGE, published on behalf of the Canadian Association of School Psychologists. Used by permission. DOI: 10.1177/0829573509331614. Originally published online (Sage OnlineFirst) February 17, 2009; version of record March 29, 2009; http://cjs.sagepub.com/content/24/1/82

Investigating the Comparability of a Self-Report Measure of Childhood Bullying Across Countries

Chiaki Konishi[a], Shelley Hymel[a], Bruno D. Zumbo[a], Zhen Li[a], Mitsuru Taki[b], Phillip Slee[c], Debra Pepler[d], Hee-og Sim[e], Wendy Craig[f], Susan Swearer[g] and Keumjoo Kwak[h]

[a]University of British Columbia
[b]National Institute for Educational Policy Research, Tokyo
[c]Flinders University
[d]York University
[e]Kunsan National University
[f]Queen's University
[g]University of Nebraska-Lincoln
[h]Seoul National University

Abstract: Responding to international concerns regarding childhood bullying and a need to identify a common bullying measure, this study examines the comparability of children's self-reports of bullying across five countries. The Pacific-Rim Bullying Measure, a self-report measure of students' experiences with six different types of bullying behavior and victimization, was administered to 1,398 grade 5 students from Australia, Canada, Japan, Korea, and United States. Multigroup confirmatory factor analysis and item response theory modeling were used to evaluate construct equivalence on the measure across different countries. Preliminary results revealed some construct differences across countries, that is, the bullying measure is measuring one construct, but that the construct is manifested differently in the different countries.

Authors' Note: The study presented herein arose out of funding support from the National Institute for Educational Policy Research in Tokyo. Portions of this research were presented at the annual meeting of the American Educational Research Association, New York, March 2008. Correspondence concerning this article should be addressed to Chiaki Konishi at the Department of Educational and Counselling Psychology, and Special Education, University of British Columbia, 2125 Main Mall, Vancouver, B.C., V6T 1Z4, Canada.

Résumé: En réponse aux inquiétudes partagées par la communauté internationale concernant la présence en enfance de formes d'intimidation et le besoin d'une mesure commune de l'intimidation, cette étude examine la comparabilité de rapports individuels d'enfants de cinq pays différents sur leurs expériences d'intimidations. The Pacific-Rim Bullying Measure, une mesure des rapports individuels d'expériences d'intimidation d'étudiants en fonction de six types de comportement d'intimidation et de victimisation, a été administrée à 1,398 étudiants de cinquième année provenant de l'Australie, du Canada, du Japon, de la Corée et des États-Unis. Une analyse factorielle multi-groupe et une modélisation théorique en fonction des réponses à des items ont été employées pour évaluer l'équivalence des concepts utilisés par cet instrument à travers les pays concernés. Les résultats préliminaires indiquent quelques variations dans les concepts d'un pays à l'autre c.-à-d., l'instrument de mesure d'intimidation mesure un concept unique, mais ce concept est manifesté de façon différente de pays en pays.

Keywords: Bullying; Comparability; Measure, cross-national study

It has been more than two decades since bullying began to attract public attention as a serious threat to the safe environment of schools. Across Europe, North America, Asia, and Australia, bullying is now recognized as a global problem (e.g., Smith et al., 1999), affecting millions of children in schools around the world. Accordingly, researchers have undertaken cross-national studies investigating differences and similarities in student reports of bullying across different countries (see Morita, Smith, Junger-Tas, Olweus, & Catalano, 1999; Smith, Cowie, Olafsson, & Liefooghe, 2002; Smith et al., 1999; Taki et al., 2006). It is important to note that these same concerns regarding cross-national studies also have implications for research and assessment practice in culturally diverse settings within a nation (e.g., Canadian schools) wherein students come from many different countries and cultures. For example, because of the large number of Korean students, one could choose to use both a Canadian and Korean version of a bullying measure in their district or school annual report or evaluation of a bullying intervention or prevention program.

The challenges facing such cross-national research are many and multifaceted. One major concern is the comparability of the measure used to assess bullying in research. That is, there is a need to choose or create a measure that taps the same underlying construct across different countries. Indeed, previous research by Smith and colleagues (Smith et al., 2002) has shown that terms used to describe "bullying" across different languages evoke different meanings regarding the type of bullying reported. Specifically, by examining children's understanding of "bullying" across 14 countries and 13 languages, Smith and colleagues (2002) demonstrated that children's understanding of the phenomenon varies considerably as a function of language and culture. Accordingly, comparative international research must evaluate "bullying" in a way that is culturally inclusive and consistent across languages.

In an effort to avoid such language-based differences, the Pacific-Rim Bullying Measure (Taki et al., 2006) was developed, asking children to report on the behaviors that are included in common definitions of bullying without reference to terms such as "bullying" that carry different meanings across languages and countries. In particular, an effort was made to define bullying consistently across countries using a description of the behavior that included reference to the three critical elements of bullying (Olweus, 1993)—intentionally, repetition, and power differential—that could be readily translated across languages. Although aggressive behavior is generally defined as any form of behavior that is intended to harm someone physically or psychologically (Baron & Richardson, 1994; Berkowitz, 1993; Olweus, 1999), bullying is regarded as a subcategory of aggressive behavior that is distinguished from general aggressive behavior in terms of its frequency of occurrence and the power imbalance between perpetrator(s) and his or her (or their) victim(s). Given all the possible sources of cross-country differences in bullying, exact matching of constructs across countries is almost certainly difficult to achieve. However, it is necessary to know whether this Pacific-Rim Bullying Measure is effective in creating a comparable self-report index of bullying across different cultural and language groups. To date, statistical evidence on the comparability of bullying measures across countries has not been examined.

The purpose of the present study was to examine whether a measurement instrument, specifically a self-report measure of bullying (i.e., the Pacific-Rim Bullying Measure; Taki et al., 2006), is comparable across different countries (i.e., Australia, Canada, Japan, Korea, and United States). Specifically, we examined whether the measure taps the same underlying latent variable and whether the construct is being measured equivalently across groups using multigroup confirmatory factor analysis (Mg-CFA) and item response theory (IRT) modeling.

In Mg-CFA, researchers are interested in finding out whether the same measurement model is invariant across samples or groups. Measurement invariance is tenable when the relations between observed variables and latent construct(s) are identical across relevant samples or groups. Indeed, Horn and McArdle (1992) contend that Mg-CFA addresses "whether or not under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute" (p. 117).

Whereas Mg-CFA is an analytical technique that evaluates construct equivalence of measures at the scale level across different groups, IRT modeling is an analytical technique that allows for an examination of construct equivalence at the item level. In IRT modeling, researchers are interested in the process underlying a person's response to a question or an item, especially the relationship between the probability of reporting a particular item response and the latent variable being measured. The present study uses Mg-CFA and IRT (in particular, differential item functioning— DIF) to evaluate the comparability of student reports of bullying and victimization (i.e., being bullied) across five countries and three languages using data obtained with the Pacific-Rim Bullying Measure.

Method

Participants

The data used in this study were collected annually in 2004 through 2006 as part of an international longitudinal project on bullying coordinated by Mitsuru Taki of the National Institute for Educational Policy Research in Tokyo, Japan. Participants included 1,398 students in fifth-grade classrooms in Australia (*n* = 130), Canada (*n* = 412), Japan (*n* = 302), Korea (*n* = 436), and United States (*n* = 118). Previous research indicates that bullying behavior is particularly evident within this age group, grade 5 to 7 age range, (Menesini et al., 1997; Morita et al., 1999; Nansel et al., 2001). Other research indicates that bullying decreases somewhat at later ages (Whitney & Smith, 1993).

Measures

Demographic information. To obtain descriptive information about the sample, participants were asked to provide information on their (a) gender, (b) birth date/age, (c) grade, and (d) ethnic background.

Bullying. The Pacific-Rim Bullying Measure (Taki et al., 2006) was used to assess students' experiences as both a bully and a victim, without relying on terms such as "bullying" that have been shown to reflect different understanding of the construct across countries and languages (Smith et al., 2002). Instead, bullying was described in behavioral terms including the three primary distinguishing characteristics of bullying as outlined by researchers (e.g., Olweus, 1993): intentionality, repetition, and power differential (see Table 1). Following a general behavioral description of such behavior, students were asked to respond to six bullying items, each reflecting a different type of bullying behavior—physical bullying, jokingly; physical bullying, on purpose; property damage; verbal bullying; social/relational bullying; and cyber/electronic bullying. A comparable set of six items tapped victimization (see Table 1). For each item, participants indicated whether the behavior occurred *never* (1), *sometimes* (2), or *once a week or more* (3)[1] reflecting how often they had taken part in (bullying) or were recipients of (victimization) each behavior with peers in the past 2 months.

Table 1
Bullying and Victimization Items

General description: Students can be very mean to one another at school. Mean and negative behavior can be especially upsetting and embarrassing when it happens over and over again, either by one person or by many different people in the group. We want to know about times when students use mean behavior and take advantage of other students who cannot defend themselves easily.

Bullying	Victimization
In the past 2 months, how often have you taken part in being mean or negative to others.	In the past 2 months, how often have other students been mean or negative to you.

1. By pushing, hitting, kicking, or other physical ways (jokingly)?
2. By pushing, hitting, kicking, or other physical ways (on purpose)?
3. By taking things from them or damaging their property?
4. By teasing, calling them names, threatening them verbally, or saying mean things to them?
5. By excluding or ignoring them, spreading rumors or saying mean things about them to others, or getting others not to like them?
6. By using computer, e-mail, or phone text messages?

Results

MG-CFA

Mg-CFA was used to evaluate four commonly investigated hypotheses for the cross-country measurement model (i.e., in our case, one factor with six observed variables), examining the bullying and victimization scales separately: (a) whether the overall structure was the same across countries/samples; (b) whether the overall structure and factor loadings were the same; (c) whether the overall structure, factor loadings, and intercepts were the same across countries/samples[2]; and (d) whether the overall structure, factor loadings, intercepts, and error variances were the same across countries/samples. Please see Wu, Li, and Zumbo (2007) for a detailed description of these four commonly investigated hypotheses and the implications thereof for research practice.

According to Cheung and Rensvold (2002) and Wu et al. (2007), the chi-square test is often too sensitive and likely to reject hypotheses that are tenable. Accordingly, these authors recommend use of the comparative fit index (CFI) test (rather than chi-square) to examine measurement invariance. If the difference in CFI values between two nested models (e.g., a model wherein the overall structure is equal vs. a model wherein the overall structure and loadings themselves are equal) is < 0.02, the more restrictive model is supported (in our example, the case wherein the overall structure and loadings were the same; Cheung & Rensvold, 2002). The Mg-CFA results of the cross-country measurement invariance are presented in Tables 2 and 3.

Table 2
Multigroup Confirmatory Factor Analysis Results for the Measurement Invariance of the Bullying and Victimization Subscales across Countries

Model	X^2	*df*	*p*	RMSEA	CFI	ΔX^2	Δdf	ΔCFI
Bullying								
1. Same overall structure	59.27	45	0.08	0.03	1.00			
2. Same overall structure and same factor loadings	125.46	65	0.00	0.06	0.99	66.37***	20	0.01
3. Same overall structure, same factor loadings, and same intercepts	400.41	85	0.00	0.12	0.94	341.14***	40	0.06
4. Same overall structure, same factor loadings, same intercepts, and same error variances	529.64	109	0.00	0.12	0.92	470.37***	64	0.08
Victimization								
1. Same overall structure	71.96	45	<0.01	0.05	0.99			
2. Same overall structure and same factor loadings	128.99	65	0.00	0.06	0.99	57.03***	20	0.00
3. Same overall structure, same factor loadings, and same intercepts	394.89	85	0.00	0.12	0.93	322.93***	40	0.06
4. Same overall structure, same factor loadings, same intercepts, and same error variances	438.27	109	0.00	0.11	0.93	366.31***	64	0.06

Note: RMSEA = root mean square error of approximation; CFI = comparative fit index; ΔCFI = difference in comparative fit indices.
***$p < 0.001$.

Table 3
Multigroup Confirmatory Factor Analysis Results for the Measurement Invariance of the Bullying and Victimization Subscales on Pairwise Comparisons

Bullying

	Japan	Australia	Korea	Canada
Australia	2			
Korea	2	4		
Canada	2	4	4	
United States	2	4	4	4

Victimization

	Japan	Australia	Korea	Canada
Australia	2			
Korea	2	2		
Canada	2	4	3	
United States	4	4	4	4

Note: 1 denotes same overall structure (not comparable); 2 denotes same overall structure and same factor loadings (not comparable); 3 denotes same overall structure, same factor loadings, and same intercepts (comparable); and 4 denotes same overall structure, same factor loadings, same intercepts, and same error variances (comparable).

Results for a variety of model-fit indices are presented in Table 2: Chi-square (X^2), CFI, and root mean square error of approximation. Following recommendations by Cheung and Rensvold (2002) and by Wu et al. (2007), CFI was used to evaluate model fit (i.e., if the difference in CFI values between two nested models is < 0.02, the more restrictive model is supported). The models supported by CFI values are highlighted in bold in Table 2. As shown in Table 2, in the case of both bullying and victimization, we can conclude that the number of factor(s) and factor loadings were the same across the five countries. However, item intercepts and error variance were not equal, implying that a student's item score may be dependent on the student's country membership, conditional on the latent variable scores (i.e., bullying or victimization).

Subsequently, pairwise comparisons between all five countries were conducted following procedures similar to those described above (i.e., if the difference in CFI values between two nested models is < 0.02, the more restrictive model is supported). Results of these pairwise comparisons are shown in Table 3. Numbers presented in Table 3 reflect the level of the model supported by the available data. Number 2 in Table 3 refers to support for a model wherein the overall structure and factor loadings themselves are found to be equal, but fails to support a model wherein the overall structure, factor loadings, and intercepts are equal. That is, the bullying measure is consistently biased against one of the countries in the planned pairs (and therefore, not comparable). Number 3 denotes support for a model wherein the overall structure, factor loadings, and intercepts are found to be equal, but failed to hold the same error variances in a planned pairwise comparison. This implies that there are either different variables operating on the measure between the countries or the same set of variables operating differently across the paired countries (Deshon, 2004). Number 4 indicates support for a model wherein the overall structure, factor loadings, intercepts, and error variances are found to be equal in a planned pairwise comparison, suggesting comparability across countries. That is, the same construct is measured, and it is measured on the same metric. Thus, if any difference in the factor score is found, there is considerable confidence that such a difference reflects results of a true difference in the amount of the measure (in this case, reported bullying) rather than a measurement artifact. We are also confident that comparing variation is meaningful regardless of group (country) membership because cross-group (country) variances are assured to be on the same metric. With regard to the bullying subscale of the Pacific-Rim Measure, all comparisons with Japan did not pass the third model test (i.e., same overall structure, same factor loadings, and same intercepts), whereas comparisons among all other countries passed the last model (i.e., same overall structure, same factor loadings, same intercepts, and same error variances). In terms of the victimization subscale, all comparisons with Japan, again, did not pass the third model test except the comparison with United States. In addition, the pair of Australia and Korea did not pass the third model test.

IRT and DIF Analyses

Information on DIF was obtained using an application of nonparametric IRT. Because of the relatively small sample size and few items in the bullying measure, nonparametric IRT was used in the present study (Ramsay, 1991). DIF is a phenomenon in which an item is found to behave differently in different subgroups, in this case, different country groups. In other words, DIF methods allow for a judgment of whether items function in the same manner for different groups of examinees, essentially flagging noncomparable items or tasks (see Zumbo, 2007; Zumbo & Hubley, 2003, for overviews of DIF). In this study, the Testgraf beta statistic was used to investigate DIF. As Zumbo and Hubley (2003) describe, Testgraf measures and displays DIF in the form of a designated area between the nonparametric item characteristic curves. This area is denoted as beta that measures the weighted expected score discrepancy between the reference group curve and the focal group curve for examinees with the same ability on a particular item. Zumbo and Witarsa (2004) proposed the following cutoff index to detect DIF in moderate-to-small-scale testing contexts (involving 500 or fewer examinees per group and, typically, less than 50 items in a scale): $|\beta| > 0.0415$ ($\alpha = 0.01$; i.e., 99th percentile of the null DIF distribution of beta). If a DIF index for a particular scale item is larger than value of 0.0415, DIF was found on the particular item.

Table 4 shows the presence of DIF for each item of the bullying measure, based on the cutoff value of 0.0415. As shown in Table 4, DIF was found for self-reports of joking physical bullying and social bullying. With respect to the victimization subscale items, in addition to the joking physical and social victimization items, results indicated DIF for items of victimization through property damage and verbal victimization. Thus, some item-level variations were observed across countries for two of the six forms for bullying and four for victimization (see Table 4).

Table 4
Differential Item Functioning (DIF) on Bullying and Victimization Items

	Is There DIF? (Composite DIF Index)	
Items	Bullying	Victimization
	In the past 2 months, how often have you taken part in being mean or negative to others.	In the past 2 months, how often have other students been mean or negative to you.
1. By pushing, hitting, kicking, or other physical ways (jokingly)?	Yes (0.076) Korea ≠ Japan; all other countries comparable	Yes (0.079) Korea ≠ United States
2. By pushing, hitting, kicking, or other physical ways (on purpose)?	No (0.019)	No (0.025)
3. By taking things from them or damaging their property	No (0.013)	Yes (0.060) Korea ≠ Australia
4. By teasing, calling them names, threatening them verbally, or saying mean things to them?	No (0.020)	Yes (0.045) Japan ≠ all other countries
5. By excluding or ignoring them, spreading rumors or saying mean things about them to others, or getting others not to like them?	Yes (0.097) Korea ≠ Japan; all other countries comparable	Yes (0.089) Korea ≠ Australia
6. By using computer, e-mail, or phone text messages?	No (0.016)	No (0.024)

Conclusions and Educational Implications of the Study

Far too often in assessment research and practice, the comparability of measures across cultural and language groups is simply assumed by fiat, if addressed at all. Our results underscore the importance of using empirical evidence to evaluate the comparability of a measurement tool to verify the meaningfulness of particular cross-cultural comparisons and use in culturally diverse contexts.

This is the first study to consider statistical evidence in examining empirically the comparability of a bullying measure across different countries and languages. In doing so, the present study highlights the importance of looking at whether a measure is tapping the same underlying construct across different groups when conducting comparative research (e.g., cross-national studies). Specifically, Mg-CFA and IRT modeling were used to test the construct stability or comparability of the bullying measure across the five different countries. Results of Mg-CFA revealed support only for a model that indicates the same overall structure and same factor loadings across five countries, suggesting that the measures tap the same dimension of bullying and victimization across countries, but in different ways. What this means is that the constructs of bullying and victimization present (or manifest) themselves in different ways for some of the countries. For example, although the factors are the same, the means and variances of the scores on these factors may be different. Furthermore, the items do not perform the same in the various countries because an item may discriminate differently in different cultures or require more of the "bullying" to equally endorse an item. Further research is needed to investigate the nature of these differences across countries. Subsequent pairwise comparisons indicated no measurement invariance between Japan and other countries on the bullying subscale, making comparisons of scores between Japanese students and students in other countries particularly suspect. Similarly, for the victimization subscale, results again suggest caution in comparing Japan and all other countries except the United States. As well, results of the victimization subscale were not comparable between Korea and Australia.

IRT analyses indicated DIF for two of the six bullying subscale items (physical bullying—jokingly and social/relational bullying, especially between Japanese and Korean students) and for four of the six victimization subscale items (physical bullying—jokingly for Korea vs. United States, bullying through property damage and social/relational bullying for Korea vs. Australia, and verbal bullying for Japan vs. all other countries). These findings suggest considerable caution in understanding simple cross-national or cross-cultural comparisons across groups based on such bullying self-report indices. They may also, however, point to further investigation into the culturally distinct meanings of a given construct. Finally, our findings highlight that although a number of researchers tend to regard Japan and Korea as the same or similar "Asian" culture, this conventional practice needs to be revisited. The distinction between these countries reminds us that culture and language are fundamental and complex.

Notes

[1] The original measure included a 4-point scale that (in English) corresponded to 1 (*never*), 2 (*sometimes*), 3 (*about once a week*), and 4 (*several times a week*). However, given variations in translations across languages, the final two response options were collapsed to ensure comparability across countries.

[2] Unequal cross-group intercepts represent the unequal scaling of factor scores with regard to the location of the latent score distribution. If the score comparison is to be on the group means of the latent variable, it is necessary to make sure that the centres of the latent variable are scaled identically across groups. This is tested by the equality in the calibration of the mean structure in addition to the variance/covariance structure (i.e., mean and covariance structure, MACS) of the observed variables.

References

Baron, R.A., and D. Richardson (1994). *Human Aggression*. New York: Plenum Press.

Berkwitz, L.B. (1993). *Aggression: Its Causes, Consequences, and Control*. New York: McGraw-Hill.

Cheung, G.W., and R.B. Rensvold (2002). Evaluating goodness-of-fit indexes for testing MI. *Structural Equation Modeling* 9: 235-255.

Deshon, R.P. (2004). Measures are not invariant across groups with error variance homogeneity. *Psychology Science* 46: 137-149.

Horn, J.L., and J.J. McArdle (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research* 18: 117-144.

Menesini, E., M. Elsea, P.K. Smith, M.L. Genta, E. Giannetti, A. Fonzi, et al. (1997). Cross-national comparison of children's attitudes toward bully/victim problems in school. *Aggressive Behavior* 23: 245-257.

Morita, Y., P.K. Smith, J. Junger-Tas, D. Olweus, and R. Catalano (1999). *Sekai No Ijime* [Bullying of the world]. Tokyo: Kaneko Shobo.

Nansel, T.R., M. Overpeck, R.S. Pilla, W.J. Ruan, B. Simons-Morton, and P. Scheidt (2001). Bullying behaviors among US youth: prevalence and association with psychosocial adjustment. *Journal of the American Medical Association* 285: 2,094-2,100.

Olweus, D. (1993). *Bullying at School: What We Know and What We Can Do*. Oxford, U.K.: Blackwell.

Olweus, D. (1999). Sweden. In: Y. Morita (editor), *Sekai No Ijime* [Bullying of the world] (pp. 90-117). Tokyo: Kaneko Shobo.

Ramsay, J.O. (1991). Kernel-smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika* 56: 611-630.

Smith, P.K., H. Cowie, R.F. Olafsson, and A.P.D. Liefooghe (2002). Definitions of bullying: A comparison of terms used, and age and gender differences, in a fourteen-country international comparison. *Child Development* 73: 1,119-1,133.

Smith, P.K., Y. Morita, J. Junger-Tas, D. Olweus, R. Catalano, and P. Slee (1999). *The Nature of School Bullying: A Cross-National Perspective*. New York: Routledge.

Taki, M., P. Slee, H. Sim, S. Hymel, and D. Pepler (July 2006). "An international study of bullying in five Pacific Rim countries." Paper presented at the biennial meeting of the International Society for the Study of Behavioral Development, Melbourne, Australia.

Whitney, I., and P.K. Smith (1993). Survey of the nature and extent of bullying in junior/middle and secondary schools. *Educational Research* 35: 3-25.

Wu, A.D., Z. Li, and B.D. Zumbo (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: a demonstration with TIMSS data. *Practical Assessment, Research & Evaluation* 12: 1-26.

Zumbo, B.D. (2007). Three generations of differential item functioning (DIF) analyses: considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly* 4: 223-233.

Zumbo, B.D., and A.M. Hubley (2003). Item bias. *In*: Rocío Fernández-Ballesteros (editor), *Encyclopedia of Psychological Assessment* (pp. 505-509). Thousand Oaks, Calif.: Sage.

Zumbo, B.D., and P.M. Witarsa (April 2004). "Nonparametric IRT methodology for detecting DIF in moderate-to-small scale measurement: operating characteristics and a comparison with the Metel Haenszel." Paper presented at the annual meeting of the American Educational Research Association, San Diego, Calif.

Chiaki Konishi is a doctoral candidate in the Department of Educational and Counselling Psychology, and Special Education at the University of British Columbia, Canada.

Dr. Shelley Hymel is a Professor of human development in education and school psychology in the Faculty of Education at the University of British Columbia, Canada.

Bruno D. Zumbo is a Professor of Measurement and Evaluation as well as Statistics at the University of British Columbia, Canada.

Zhen Li is a doctoral candidate in the Department of Educational and Counselling Psychology, and Special Education at the University of British Columbia, Canada.

Mitsuru Taki is a Senior Researcher in the National Institute for Educational Policy Research, Tokyo, Japan.

Phillip Slee is a Professor in the School of Education at the Flinders University, Australia.

Debra Pepler is a Professor in the Department of Psychology at York University.

Hee-og Sim is an Associate Professor in the Department of Home Management at Kunsan National University, Korea.

Wendy Craig is a Professor in the Department of Psychology at Queen's University, Canada.

Susan Swearer is an Associate Professor in the Department of Educational Psychology at the University of Nebraska-Lincoln, U.S.A.

Keumjoo Kwak is a Professor in the Department of Psychology at Seoul National University, Korea.