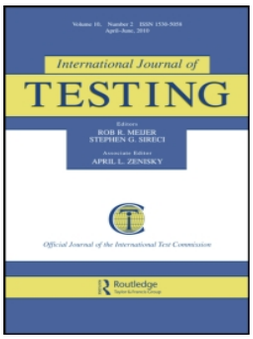

International Journal of Testing

ISSN: 1530-5058 (Print) 1532-7574 (Online) Journal homepage: www.tandfonline.com/journals/hijt20

Testing Practices and Attitudes Toward Tests and Testing: An International Survey

Arne Evers, Carina M. McCormick, Leslie R. Hawley, José Muñiz, Giulia Balboni, Dave Bartram, Dusica Boben, Jens Egeland, Karma El-Hassan, José R. Fernández-Hermida, Saul Fine, Örjan Frans, Grazina Gintilienė, Carmen Hagemeister, Peter Halama, Dragos Iliescu, Aleksandra Jaworowska, Paul Jiménez, Marina Manthouli, Krunoslav Matesic, Lars Michaelsen, Andrew Mogaji, James Morley-Kirk, Sándor Rózsa, Lorraine Rowlands, Mark Schittekatte, H. Canan Sümer, Tono Suwartono, Tomáš Urbánek, Solange Wechsler, Tamara Zelenevska, Svetoslav Zanev & Jianxin Zhang

To cite this article: Arne Evers, Carina M. McCormick, Leslie R. Hawley, José Muñiz, Giulia Balboni, Dave Bartram, Dusica Boben, Jens Egeland, Karma El-Hassan, José R. Fernández-Hermida, Saul Fine, Örjan Frans, Grazina Gintilienė, Carmen Hagemeister, Peter Halama, Dragos Iliescu, Aleksandra Jaworowska, Paul Jiménez, Marina Manthouli, Krunoslav Matesic, Lars Michaelsen, Andrew Mogaji, James Morley-Kirk, Sándor Rózsa, Lorraine Rowlands, Mark Schittekatte, H. Canan Sümer, Tono Suwartono, Tomáš Urbánek, Solange Wechsler, Tamara Zelenevska, Svetoslav Zanev & Jianxin Zhang (2017) Testing Practices and Attitudes Toward Tests and Testing: An International Survey, International Journal of Testing, 17:2, 158-190, DOI: 10.1080/15305058.2016.1216434

To link to this article: https://doi.org/10.1080/15305058.2016.1216434

Published online: 03 Sep 2016.

International Journal of Testing, 17: 158–190, 2017
Published with license by Taylor & Francis
ISSN: 1530-5058 print / 1532-7574 online
DOI: 10.1080/15305058.2016.1216434

Routledge
Taylor & Francis Group

Testing Practices and Attitudes Toward Tests and Testing: An International Survey

Arne Evers
Unit Work & Organizational Psychology, University of Amsterdam, the Netherlands

Carina M. McCormick
Buros Center for Testing, Nebraska, USA

Leslie R. Hawley
Nebraska Academy for Methodology, Analytics and Psychometrics, USA

José Muñiz
Department of Psychology, University of Oviedo, Spain

Giulia Balboni
Department of Philosophy, Social and Human Sciences and Education, University of Perugia, Italy

Dave Bartram
University of Pretoria, South Africa

Dusica Boben,
Drustvo Psihologov Slovenije, Slovenia

Jens Egeland
Vestfold Hospital Trust, Norway

Karma El-Hassan
Office of Institutional Research & Assessment, American University of Beirut, Lebanon

José R. Fernández-Hermida
Spanish Psychological Association, Spain

Saul Fine
Midot, Ltd. & Department of Psychology, University of Haifa, Israel

Örjan Frans
Department of Psychology, University of Uppsala, Sweden

Grazina Gintiliené
Department of General Psychology, Vilnius University, Lithuania

Carmen Hagemeister
Department of Psychology, Technical University Dresden, Germany

Peter Halama
Department of Psychology, University of Trnava, Slovakia

Dragos Iliescu
Department of Psychology, University of Bucharest, Romania

Aleksandra Jaworowska
Psychological Test Laboratory of the Polish Psychological Association, Poland

Paul Jiménez
Department of Psychology, University of Graz, Austria

Marina Manthouli
Association of Greek Psychologists, Greece

Krunoslav Matesic
Department of Psychology, University of Zagreb, Croatia

Lars Michaelsen
Danish Psychological Association, Denmark

Andrew Mogaji
Department of Psychology, Benue State University, Nigeria

James Morley-Kirk
China Select, China

Sándor Rózsa
Department of Personality and Health Psychology, Eotvos Lorand University, Hungary

Lorraine Rowlands
New Zealand Council for Educational Research, New Zealand

Mark Schittekatte
Faculty of Psychology and Educational Sciences, University of Gent, Belgium

H. Canan Sümer
Department of Psychology, Middle East Technical University, Turkey

Tono Suwartono
Faculty of Teacher Training, Muhammadiyah University of Purwokerto, Indonesia

Tomáš Urbánek
Institute of Psychology, Academy of Sciences, Czech Republic

Solange Wechsler
Center for Theology and Human Sciences, Pontifical Catholic University of Campinas, Brazil

Tamara Zelenevska
Latvian Professional Psychologist Association, Latvia

Svetoslav Zanev
Oganizzazioni Speciali, Bulgaria

Jianxin Zhang
Chinese Academy of Sciences, China

Correspondence should be sent to Arne Evers, Work & Organizational Psychology, University of Amsterdam, Nieuwe Achtergracht 129 B, Room 2.10, 1018 WS Amsterdam, the Netherlands.
E-mail: a.v.a.m.evers@uva.nl

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hijt.

On behalf of the International Test Commission and the European Federation of Psychologists' Associations a world-wide survey on the opinions of professional psychologists on testing practices was carried out. The main objective of this study was to collect data for a better understanding of the state of psychological testing worldwide. These data could guide the actions and measures taken by ITC, EFPA, and other stakeholders. A questionnaire was administered to 20,467 professional psychologists from 29 countries. Five scales were constructed relating to: concern over incorrect test use, regulations on tests and testing, internet and computerized testing, appreciation of tests, and knowledge and training relating to test use. Equivalence across countries was evaluated using the alignment method, four scales demonstrated acceptable levels of invariance. Multilevel analysis was used to determine how scores were related to age, gender, and specialization, as well as how scores varied between countries. Although the results show a high appreciation of tests in general, the appreciation of internet and computerized testing is much lower. These scales show low variability over countries, whereas differences between countries on the other reported scales are much greater. This implies the need for some overarching improvements as well as country-specific actions.

Keywords: psychological testing, testing practices, test use, International Test Commission, European Federation of Psychologists' Associations

Many countries intensively use educational and psychological tests (e.g., Bartram & Coyne, 1999; Evers, Zaal, & Evers, 2001; Fine, 2013; Muñiz, Prieto, Almeida, & Bartram, 1999). Sensible test use requires both that the test demonstrate adequate psychometric properties and that the results are used appropriately. In addition, appropriate test use also requires that the test user verifies and evaluates if adequate reliability and validity evidence is available for the intended test interpretation for a specified use, and if not, to provide this information himself or herself (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Information about the psychometric quality of tests is becoming available in an increasing number of countries (Evers, 2012). Because tests are important tools with significant consequences for the persons tested and organizations using the scores, it is also of interest to know the attitudes of psychologists with respect to tests and the ways tests are used. Therefore, the European Federation of Psychologists' Associations (EFPA) initially took the initiative to investigate psychologists' attitudes toward various aspects of testing in 2000 (Muñiz et al., 2001).

The 2000 Survey Administration

In the 2000 administration, a comprehensive survey was conducted in six European countries (Belgium, Croatia, the Netherlands, Slovenia, Spain, and the United Kingdom). Factor analysis revealed five attitude factors: (concern over)

incorrect test use, regulations on tests and testing, appreciation of tests, knowledge and training, and permissiveness or qualifications for test use. The results showed that in general European psychologists had a positive attitude toward tests and testing. Their scores also indicated a desire for greater involvement of the professional organizations in the regulation of tests as well as more information on technical aspects of tests. Finally, results clearly indicated a demand for ongoing training, because training provided up to first-degree level was found to be insufficient (Muñiz et al., 2001). This survey resulted in valuable information contributing to the projects and actions carried out by EFPA and the International Test Commission (ITC) to improve psychologists' use of tests. Muñiz and Bartram (2007) provided an overview of these projects. More recent projects are the European actions on test-user certification (Bartram, 2011) and the international norms for assessment procedures (International Organization for Standardization, 2011a, 2011b).

The 2009 Survey Administration

Almost 10 years later, in 2009, the EFPA Standing Committee on Tests and Testing (EFPA-SCTT) considered it appropriate to reassess European psychologists' opinions of tests (Evers et al., 2012). Seventeen European countries participated in the 2009 survey, including the six countries participating in the first survey. Direct comparison of the 2000 and 2009 survey results on scale level was not possible because of the changes made in the questionnaire (as detailed in the Questionnaire subsection in this article). For similar items, the correlations between the means in the 2000 and 2009 survey were computed to have some measure of stability for test attitudes in the six countries surveyed in 2000.

Prior work compared the mean item responses for these countries between the two administrations and documented key changes over that timespan (Evers et al., 2012). The correlations between the same items over time ranged from 0.05 to 0.94, with a median value of 0.75. The value of 0.05 was an outlier, however, and concerned the item "I use tests regularly," which is not an attitude item. The median correlation of 0.75 shows that the pattern of test attitudes for the included countries over time was rather stable but also that there was some change.

Items with the greatest differences in overall means (combining the six countries) between 2000 and 2009 were identified and further examined. Compared to 2000, psychologists in 2009 demonstrated less concern over need for enforceable test quality standards ($d = 0.50$) but more concern about illegal copying of test materials ($d = 0.23$). Psychologists in 2009 were more satisfied with the sufficiency of information about test quality ($d = 0.33$). Differences on the other items over time were smaller or not significant in the total group, although some differences were more pronounced within individual countries. For example, whereas in the total group the effect size was near zero ($d = 0.03$) for an item

addressing whether interpretation and feedback of test results should be restricted to psychologists, Slovenian psychologists indicated a substantially more liberal attitude in 2009 than in 2000 ($d = 0.65$).

It is also of interest that compared to 2000, participating psychologists in 2009 felt better equipped for test use after completing their masters' degrees. However, they reported that their knowledge was based more on training after receiving their masters' degrees than during the degree program itself. This increase in psychologists' preparations for test use corresponds chronologically to higher investments in education for psychologists and increased availability of information for test users. A more comprehensive description of the results at the item level is given in Evers and colleagues' work (2012).

Besides performing a follow-up of the results obtained in the first survey, important reasons for this reassessment were to broaden the number of included countries and to assess the opinions of psychologists regarding technical advances that have emerged over the past decade in the field of testing. These advances have a noticeable impact on the way psychologists practice their profession in general and in particular on the use of tests (Bartram & Hambleton, 2006). Therefore, in the 2009 survey various questions were added with respect to computer-based and Internet testing. Items were added addressing the developments in computer-based testing in general and more specifically to the testing via the Internet as well as related issues of unproctored (or unsupervised) testing and computer-generated feedback.

An important observation obtained from the 2009 data was that the differences between the countries on one scale (appreciation of tests) were small, but that the differences on the other four scales used (concerns over incorrect test use, regulations on tests and testing, computer-based and Internet testing, knowledge and training) showed large effect sizes (Cohen, 1988), with differences between the extreme scoring countries of 1.0 to 1.5 *d*. Another important finding was that although the participating psychologists showed a positive attitude toward the use of psychological tests in general, they showed a relatively low level of appreciation of Internet or computer-based testing.

Expansion of the 2009 Survey

Presented with the results from the 2009 administration, the board of the ITC asked the EFPA Board of Assessment to expand the survey to countries outside Europe. (The name of the SCTT has since been changed into Board of Assessment.) First, it was relevant to investigate whether the opinions about tests in countries worldwide showed means and variations that differed from countries in Europe. In addition, a better understanding of the state of psychological testing worldwide could guide the actions and measures taken by ITC and other stakeholders.

For this expansion of the survey four research questions were formulated. First, overall, how do psychologists internationally view the current state of

testing practices and test use, including training and computerized administration or reporting? Second, how do these views vary across gender, age, and specialization? Third, how do these views vary across countries? Which countries have average ratings significantly higher or lower than the mean? Fourth, what policy shifts will be suggested by psychologists' views on testing as reported in the survey? How will these suggested goals differ across countries?

As a consequence, the objective of the present study was to administer the survey used in 2009 to psychologists in as many countries as possible in all parts of the world. In 2012, data from 12 additional countries were gathered, yielding a 29 countries for which data are available. The combined data from the 2009 administration and the 12 additional countries were analyzed and are reported in this article. Expanding on the work of Evers and colleagues (2012) in which analysis of variance was used to investigate differences between the original 17 countries, the current study used multilevel modeling techniques to account for different sources of variation between countries at multiple hierarchical levels.

METHOD

The Questionnaire

The questionnaire used for the 2009–2012 administration (EFPA Questionnaire on Test Attitudes of Psychologists—EQTAP, see Appendix) was based on, but not identical to, the one used in 2000 (which in turn was partly based on the work of Eyde et al., 1988, 1993). The main difference was that due to developments in the field of testing, six items about the attitude toward diverse aspects of computer-based testing and Internet-delivered testing were added, as described previously. Because of these additional items, the questionnaire became too long, and therefore, four items showing low or unstable factor loadings in the 2000 survey (Muñiz et al., 2001) were deleted.

Minor changes in three other items were made to clarify or update the formulations. In order to make the survey applicable for countries outside Europe, reference to EFPA was broadened to ITC/EFPA in two items. All 32 attitude items were administered on 5-point Likert-type scales (see Appendix). Further, the questionnaire contained an open-ended question asking respondents to list the three tests they use most frequently, and three questions concerning biographical information (age, gender, and field of specialization).

The items were originally formulated in English. Within each country, the national representatives were responsible for organizing the translation into the country's language or determining to administer the survey in English. In six countries the survey was not translated, but administered in English (Greece, New Zealand, Nigeria, Norway, Sweden, and the United Kingdom). In two countries the survey was offered in both the English and the local language

version (Latvia and Lebanon). Eight countries used a translation–back-translation procedure as recommended by ITC (Hambleton, Merenda, & Spielberger, 2005) (Brazil, China, Israel, Lebanon, Lithuania, Romania, Turkey, and Spain). For the remaining countries, the methods used to translate and evaluate the survey varied, including methods such as independent translation and reconciliation by two or more bilingual psychologists and evaluation of the translation by all members of the national committee on testing.

Survey Administration

For the 2009 survey, the national representatives in the EFPA-SCTT were asked to participate and to organize the survey in their respective countries. For the additional global data gathering, an e-mail was sent to all "friends of the ITC" with the same request. Friends of the ITC are individual members of the ITC, known to be active in some way in their respective countries (about 200 in total).

The process of distribution and administration of the questionnaire varied.[1] Distribution and administration details can be found in Table 1. Most of the countries that invited participants personally sent a reminder within some weeks after the first e-mail.

Participants

The total sample consisted of 20,467 psychologists who answered at least 24 out of the 32 attitude items. Setting the limit at eight missing items allowed for retaining the data of about 300 respondents who did not answer item 25 only. (Item 25 includes eight subquestions; see the Appendix for the text of item 25.) For the total group, the response rate is 11.3% (see Table 2). The response rates vary from 3.4% (Germany) to 42.2% (Slovakia). The variation in response rates may be caused by the variety of methods used for approaching respondents. The combination of computer and paper-and-pencil administration and the more personal approach in Slovakia resulted in the highest response rates.

For 15 countries, the size of the populations, as given in Table 2, is equal to the number of members of the psychological associations. Exceptions are Austria, Brazil, China, Germany, Hungary, Indonesia, Israel, Italy, Latvia, Nigeria, Slovakia, Slovenia, and the United Kingdom. In Germany and Hungary, only half of the members were approached; in Italy, about 25%. In the United

[1]We evaluated the possibility of a systematic effect of administration method on scores. Because administration methods were consistent within most countries, some differences might be attributable to differences between the samples in the countries. Therefore, this effect was tested as a fixed effect in the multilevel models described in the Methods and Results section. Administration method did not have a significant fixed effect for any of the five scales, providing evidence that there was not a systematic influence of administration method on scores.

TABLE 1
Sampling Method

Administration		
Paper & pencil	Brazil, Croatia, Greece, Hungary, Indonesia, Nigeria, Spain	
Computer	Austria, Belgium, Bulgaria, China, Czech Republic, Denmark, Germany, Italy, Israel, Latvia, Lebanon, Lithuania, the Netherlands, New Zealand, Norway, Romania, Slovenia, Sweden, Turkey, United Kingdom	
Both paper & pencil and computer	Poland, Slovakia	
Distribution		
Probability sampling		
All members or a random sample of the members of the national psychological association are invited personally	By post	Croatia, Spain
	By e-mail	Bulgaria, Czech Republic, Denmark, Italy, Israel, Latvia, Lithuania, the Netherlands, Norway, Romania, Slovenia, Sweden, United Kingdom
Nonprobability sampling		
Invitation on website or in newsletter of national psychological association or in other media	Austria, Belgium, Germany, Lebanon, New Zealand, Poland, Slovakia, Turkey	
Handing out at psychological conferences, post-academic courses, or general meetings	Brazil, Greece, Hungary	
Snowball method	Indonesia, Nigeria	
Invitation of a selected group	China	

Notes. In addition in Israel, Latvia, and Slovenia, an invitation to participate was sent to respectively the Psychometric Association, the Association of Organizational Psychologists, or the Chamber of Clinical Psychologists. In Slovakia a message was sent to the mobile phones of the members of the Slovak Psychological Chamber. In China only the members of the Division of Psychological Measurement were contacted.

Snowball or chain-referral sampling is a nonprobability sampling technique. The first step is to identify initial subjects who are known members of the population. These subjects recruit future subjects from among their acquaintances (chain referral), and so on.

Kingdom, the invitation was sent to Chartered Psychologists only. In Austria, Israel, Latvia, Slovakia, and Slovenia, apart from the members of the psychological association, an invitation to participate was also sent to the members of chambers or associations in specific psychological areas. In Brazil, China, Greece, Indonesia, and Nigeria, only selective parts of the population were approached, which may have resulted in an overrepresentation of psychologists working in specific professional fields and/or psychologists who engage in conferences or activities of the national psychological association. For these latter

TABLE 2
Population Size and Number of Respondents per Country

Country	Population *N*	Sample *N*	Response %
Austria	3891	529	13.6%
Belgium	3000	423	14.2%
Brazil	400	70	17.5%
Bulgaria	1000	199	19.9%
China	450	178	39.6%
Croatia	1700	327	19.2%
Czech Republic	3800	271	7.1%
Denmark	4345	1189	27.4%
Germany	6500	222	3.4%
Greece	1000	86	8.6%
Hungary	1046	114	10.9%
Indonesia	200	48	24.0%
Israel	2796	350	12.5%
Italy	23,000	5482	23.8%
Latvia	150	33	22.0%
Lebanon	115	20	17.4%
Lithuania	275	107	38.9%
Netherlands	12,262	1984	16.2%
New Zealand	2165	75	3.5%
Nigeria	—	103	—
Norway	6246	942	15.1%
Poland	10,000	527	5.3%
Romania	20,000	1795	9.0%
Slovakia	600	253	42.2%
Slovenia	515	128	24.9%
Spain	51,545	3077	6.0%
Sweden	7037	848	12.1%
Turkey	1224	293	23.9%
United Kingdom	16,228	794	4.9%
Total	181,490	20,467	11.3%

14 countries, the population numbers as given in Table 1 are equal to the numbers who were actually invited to participate or could have had access to the invitation. Although these numbers are rather precise (or in some countries rounded off to the nearest hundred), they overestimate the population in some countries, for example, when newsletters were used as a way of invitation to participate, because not all psychologists read the newsletters. As a consequence, the response rates may be underestimated.

Particularly because of the low response rates, it was important to check the representativeness of the samples on some background variables. Therefore, the national associations of the 29 countries were asked to supply information on

gender, age, and professional field of their members. The comparison values of these variables for the sample and the population are given in Table 3. The associations in Austria, Brazil, Croatia, Greece, the Netherlands, New Zealand, Norway, Spain, and Turkey were able to provide precise information for all three background variables. In most cases where no precise information was available, associations provided best estimates. (See the starred figures in Table 3.) However, some countries did not or could not provide precise information or best estimates on one or more variables.

TABLE 3
Demographic Characteristics of Sample and Population

	% Female		Mean Age		% Clinical		% Education		% Work		% Other	
Country	Sample	Pop.	Sample	Pop.	Sample	Pop.	Sample	Pop.	Sample	Pop.	Sample	Pop.
Austria	84.3	83.8	37.50	41.01	52.0	51.7	19.0	21.8	18.0	18.6	11.0	7.9
Belgium*	75.4	75.0	37.53	40.00	42.3	40.0	40.2	30.0	5.9	20.0	11.6	10.0
Brazil	88.1	89.0	36.20	35.00	34.7	52.0	12.2	12.0	22.4	12.0	30.6	24.0
Bulgaria*	80.9	75.0	39.01	—	13.6	40.0	22.1	40.0	48.2	15.0	16.1	5.0
China*	62.7	70.0	32.46	—	21.3	—	41.0	—	28.7	—	9.0	—
Croatia	88.1	88.0	37.76	37.76	26.0	32.0	35.2	32.0	14.7	12.0	24.2	24.0
Czech Republic*	71.6	74.0	41.25	40.00	40.3	40.0	26.2	25.0	16.0	25.0	17.5	10.0
Denmark	75.9	77.1	46.66	46.70	56.0	—	26.2	—	9.5	—	8.3	—
Germany*	53.4	65.0	42.47	46.00	45.5	65.0	11.4	4.0	28.2	20.0	15.0	11.0
Greece	97.6	70.5	35.47	40.00	82.9	62.7	17.1	28.1	0.0	7.1	0.0	2.1
Hungary	90.4	77.2	36.54	—	65.8	59.4	24.6	17.8	8.8	9.8	0.9	13.1
Indonesia	69.8	—	36.23	—	8.3	—	85.4	—	6.2	—	0.0	—
Israel	74.9	70.0	51.79	—	36.9	50.0	47.4	44.0	11.1	6.0	4.6	0.0
Italy*	80.2	80.0	38.82	42.00	72.0	65.0	8.1	17.0	6.3	10.0	13.6	8.0
Latvia*	84.8	91.8	39.65	40.00	24.2	9.0	42.4	46.0	21.1	9.0	12.1	36.0
Lebanon	80.0	80.0	35.30	—	60.0	—	35.0	—	0.0	—	5.0	—
Lithuania	91.6	85.5	36.85	—	29.9	36.7	55.1	31.8	2.8	19.8	12.1	11.7
Netherlands	72.9	74.3	44.90	46.75	63.3	53.2	7.4	21.4	19.8	19.3	9.6	6.2
New Zealand	84.0	70.6	42.91	47.42	85.3	61.6	5.3	7.9	6.7	7.5	2.7	23.0
Nigeria	28.0	—	42.50	—	34.1	—	3.3	—	24.2	—	22.0	—
Norway	58.1	64.0	42.17	56.00	83.8	73.4	5.3	8.2	4.1	5.0	6.8	13.4
Poland*	90.1	75.0	39.65	37.50	35.0	25.0	50.8	30.0	9.5	15.0	4.7	30.0
Romania*	83.1	80.0	38.95	37.50	29.9	35.0	27.1	17.1	43.1	47.9	0.0	0.0
Slovakia*	80.2	75.0	42.13	40.00	38.2	40.0	40.6	30.0	14.9	20.0	6.4	10.0
Slovenia*	77.3	75.0	40.45	43.00	41.4	30.0	22.7	35.0	18.8	25.0	17.2	10.0
Spain	71.4	78.1	41.83	40.58	64.2	68.4	17.6	15.3	7.7	8.1	10.6	8.2
Sweden*	68.2	72.0	49.12	—	68.0	50.0	12.3	30.0	12.4	20.0	7.3	0.0
Turkey	72.7	77.5	34.47	37.72	56.0	28.1	17.7	15.5	6.1	4.3	20.1	52.1
United Kingdom	62.5	—	47.77	—	45.1	45.4	20.7	13.2	23.2	15.8	11.1	25.6
Total	75.6	76.6	41.39	41.74	58.0	54.8	17.8	18.2	13.9	16.5	10.2	10.5

Note. *Indicates population values on gender, age, and/or professional field are estimates.

Data Analyses

First, analyses related to sample demographics were conducted using IBM SPSS Statistics version 22. Chi-square tests were used to examine the representativeness of the total sample, compared to the population. Descriptive statistics of the demographic data provided by the participants, broken down by country, were also calculated in this step.

Dimensionality of the survey was assessed using factor analysis, the results of which guided scoring. Exploratory factor analysis was completed using the software program M*plus* version 7.3 (Muthén & Muthén, 1998–2012). The country-level clustering of the data was accounted for using the program's feature "TYPE = COMPLEX," and estimation was completed using maximum likelihood with robust standard errors. Due to the expected association between factors, an oblique rotation method was selected over an orthogonal method. Selection of the number of factors proceeded iteratively with removal of problematic items and was based on a comparison of Root Mean Square Error of Approximation (RMSEA) values for the different solutions, the eigenvalues for each factor, and the pattern of loadings in the various solutions. The number of and configuration of factors supported by this analysis were used in the remaining analyses reported in the article. More details about how items were selected for removal and how the numbers of factors were selected are provided in the Results section.

Next, score calculations and psychometric analyses were completed using IBM SPSS Statistics version 22. Scores for each of the scales were created by calculating the average of responses to items corresponding to each scale. Classical test theory reliability for each scale was calculated as coefficient alpha. Descriptive statistics of these scores were then calculated for each country. All references to scale scores in the article refer to these scale means rather than any factor scores that could have been calculated. Overall psychometric properties of each scale were evaluated, and psychometric properties of the scales for the individual countries were also considered, as described next.

This study used a relatively new method called alignment to evaluate comparability of the scales between countries. Common practice dictates that in order for factor means to be comparable between groups, factor loadings and intercepts need to be invariant for these groups (Millsap, 2011). Invariance in loadings is typically referred to as metric invariance, and invariance in intercepts is typically called scalar invariance. In addition to general critiques of this conservative invariance standard, this level of invariance may be especially impractical when comparing many groups. For example, "with many groups, the usual multiple-group CFA approach is too cumbersome to be practical due to the many possible violations of invariance" (Asparouhov & Muthén, 2014, p. 1).

The alignment method was specifically developed to overcome problems that arise from attempting commonly used invariance testing methods with many

groups (Muthén & Asparouhov, 2013; Asparouhov & Muthén, 2014). Compared to the traditional sequential method, the alignment method does not assume measurement invariance; instead, it estimates all groups' factor means and variances while the program discovers the optimal invariance pattern utilizing a simplicity function, similar to rotation methods used in exploratory factor analysis. In calculating the results, the alignment optimization seeks a solution that minimizes the overall degree of noninvariance, and some instances of noninvariance are generally expected due to this method of optimization. Countries showing limited noninvariance are not removed from the analysis because "measurement invariance studies benefit from information on which groups contribute to noninvariance. This information is readily obtained by the alignment method" (Muthén & Asparouhov, 2013, p. 30).

Analyses for the alignment method of evaluating measurement invariance across groups were completed in M*plus* 7.3 using robust maximum likelihood estimation. This procedure was repeated separately for each scale, and results for each item within the corresponding scale were examined. For most options, the default settings were used, but the models were defined as fixed rather than free based on preliminary analyses. The method produces detailed results that indicate whether approximate invariance was achieved for each group and each item. (See Asparouhov and Muthén, 2014, for more details on how results are presented and their interpretation.)

Scale scores were analyzed with multilevel modelling in SAS version 9.3 (SAS Institute Inc., 2011) in order to account for the clustering of respondents within countries. Using the PROC MIXED procedure, a series of multilevel regression analyses was completed, using a separate but parallel model for each of the scales. The results of these analyses indicated whether gender, age, and specialization were associated with significant differences in the scale scores, and to what magnitude these characteristics were associated with differences in scale scores.

Predictors for gender and specialization were dummy coded into dichotomous variables with female and clinical specialization as the reference variables. These categories were selected for use as the reference groups because they were the most prevalent in the sample. The effect estimates for the comparison of male, as compared to female, and for educational or work specializations, as compared to clinical specialization, were included as fixed effects with significance tests. Age was centered at 40 years, near the average age for the sample. The interaction between gender and specialization was evaluated by testing an interaction between gender and each of the two-dummy coded variables of educational and work specializations. Thus, there were a total of 10 interactions tested. In all but one of these tests, the interaction was not significant; for reasons of parsimony these interactions were therefore not included for any of the scales.

In addition, random intercepts for each country and their associated confidence intervals were evaluated to determine which countries had average scores

on each scale that were significantly higher or lower than the overall mean. For each scale, countries were then grouped into those with high scores, those with low scores, and those with scores not significantly different than the mean, after accounting for the country's own demographic makeup. This analysis was repeated for the overall ratings as well as after accounting for country differences in gender, age, and specialization. More detail about this method is provided with the corresponding results. For the results of the between-country multilevel model comparisons, only four scales are included due to concerns with equivalence across countries for the final scale.

RESULTS

Sample Representativeness and Demographic Characteristics

As can be observed in Table 3, the predominance of women among professional psychologists is clear, being about 75% in the total sample (as well as in the total population). The same applies to respondents working in the clinical area (58.0% in the total sample, 54.7% in the population). The mean age in the sample is 41.4 years (41.7 in the population). However, it can also be observed that there is considerable variation between countries in composition of sample and population with respect to these three demographic variables.

Results from the initial Chi-square tests in SPSS showed significant differences between the composition of the sample and the population for gender ($\chi^2 = 31.00$, df $= 1$, $p < 0.0001$) as well as field of specialization ($\chi^2 = 121.48$, df $= 3$, $p < 0.0001$). However, these significant results may be attributable to the large sample size because the effect sizes of the differences are very small ($w = 0.02$ and 0.08, for gender and field of specialization, respectively; see Cohen, 1988). The representativeness of the sample with respect to age could not be tested, because standard deviations for the population are missing, but the absolute difference of 0.35 years in mean age between sample and population can also be considered very small.

Exploratory Factor Analysis

First, the number of distinct factors was evaluated in multiple ways, converging on a five-factor solution. The five-factor solution was the first for which the RMSEA was less than 0.05, so we especially considered the solutions for four, five, and six factors. Visual inspection of the scree plot (Figure 1) provided some support for a five-factor solution, based on change between factors five and six (1.63 and 1.20, respectively). Factors six and beyond demonstrate a relatively smooth line of lower eigenvalues, while factors four and five have eigenvalues similar to each other (1.72 and 1.63, respectively). In addition, the

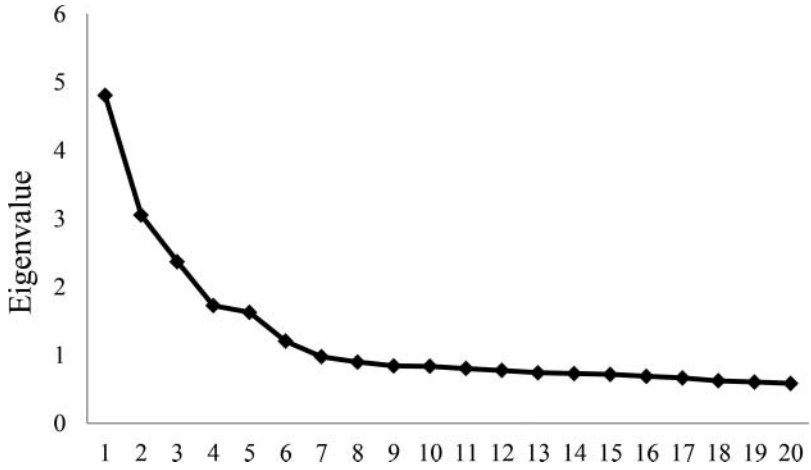

FIGURE 1
Scree plot of exploratory factor analysis.

five-factor solution showed the clearest pattern of moderately high loadings for each item on exactly one factor. The pattern of item loadings also had substantive support—for items whose highest loading was on a given factor, the content of those items was similar.

Some preliminary item analysis, including examination of the item wording, suggested that items four, five, and nine might be most problematic. (Readers can view the full text of these items in the Appendix and notice possible problems with their wording.) In evaluating item loadings, we particularly examined items that had loadings substantially below 0.3. We iteratively evaluated problematic items along with the optimal number of factors to protect against poorly performing items from contaminating decisions about the number of factors. However, the five-factor solution was preferred before and after the removal of the most problematic items. Across each of the factor solutions for three to seven factors, items four and nine consistently showed poor patterns of loadings and were removed. Specifically, these two items produced low absolute values for the rotated factor loading (<0.3) and often had loadings with opposite signs of the other items associated most closely with that factor, even for models with a different number of factors.

The solution for factor three (which contains items relating to computer and Internet testing) was more nuanced, requiring further examination using confirmatory factor analysis with M*plus*. To determine how to improve this factor, the residual variances of each item were examined, along with the unstandardized factor loadings. Item five had the largest residual variance and once it was removed, all remaining items' loadings increased and each had their residual variances reduced. This question likely operated differently because it does not

TABLE 4
Scale and Item Statistics: Factor Loadings and Alpha Reliability

Scale 1		Scale 2		Scale 3		Scale 4		Scale 5	
Item	Loading	Item	Loading	Item	Loading	Item	Loading	Item	Loading
25a	0.50	3	0.50	7	0.73	21	0.56	1	0.75
25b	0.74	8	0.44	10	−0.33	22	0.83	2	0.60
25c	0.71	11	0.55	13	−0.38	23	0.88	6	0.37
25d	0.69	12	0.66	15	0.73	24	0.32		
25e	0.75	14	−0.27	17	−0.41				
25f	0.66	16	−0.31	20	−0.40				
25g	0.77	18	−0.24						
25h	0.75	19	0.58						
Total		Total		Total		Total		Total	
Alpha	0.88	Alpha	0.66	Alpha	0.68	Alpha	0.70	Alpha	0.56

clearly ask about attitudes toward testing but rather about more objective observations. Final rotated factor loadings for the items with the strongest loadings for each factor are provided in Table 4 along with coefficient alpha reliability for the overall sample.

Next, factor solutions were compared to the results from the 2000 administration, which included fewer items than the current administration. Factors one, four, and five in the current study are similar to the factors labelled "Incorrect test use," "Appreciation of tests," and "Knowledge and training" from analyses of the 2000 administration, respectively. Factor two can be characterized as a combination of the factors "Regulations on tests and testing" and "Qualifications for test use" found in the 2000 survey and will be labelled "Regulations on tests and testing" for the current administration. Factor three is new and contains items concerning the attitude toward Internet and computerized testing. After recoding negatively worded items, the mean scores on the five scales, as well as the correlations between the scales,[2] were computed. In general, the correlations between the scales were low, the highest correlation being 0.19 between the scales concerns over incorrect test use and regulations on tests and testing (see Table 5). The low correlations between scales support the argument that the scales each operate distinctly.

For scale 1, Concerns over incorrect test use, higher scores indicate more concern that test use problems occur. For scale 2, Regulations on tests and testing, higher scores indicate more stringent views favoring regulations on tests and testing. For scale 3, Internet testing, higher scores indicate more positive beliefs about the value and effectiveness of Internet-based testing and computer-

[2]Throughout the article, reported scores are the mean scores for each scale rather than factor scores.

TABLE 5
Correlations Between Total Scores for Each Scale

Scale	Scale 2	Scale 3	Scale 4	Scale 5
1. Concerns over incorrect test use	0.19	−0.07	−0.13	−0.11
2. Regulations on tests and testing		−0.18	0.16	−0.11
3. Internet testing			0.02	0.08
4. Appreciation of tests				−0.05
5. Knowledge and training				

generated score reports. For scale 4, Appreciation of tests, higher scores indicate belief that tests are valuable to the respondent psychologist as well as the field of psychology in general. For scale 5, Knowledge and training, higher scores endorse the sufficiency of psychological training within educational experiences in preparing practitioners for appropriate test use.

Invariance Testing

Results from the alignment method of evaluating invariance generally supported use of the scales in the different countries, except for scale 5. Initial attempts at analysis revealed problems with item variance for Romania with scale 1, so Romania was omitted from this scale only in the alignment analysis. Unsurprisingly, a large number of issues interfered with the analysis for scale 5, which we have described as being untrustworthy throughout this Results section, and alignment results for scale 5 are not presented. We have removed this scale also from all further analyses except when results on item level are reported.

Results for each item were then evaluated, specifically the countries that demonstrated invariance in loadings or intercepts. (These tests correspond to metric invariance and scalar invariance, respectively.) For scales 2 and 4, all items produced approximate invariance in loadings for at least 85% of countries. Scales 1 and 3 had even more favorable results, demonstrating invariance in loadings for all country-item combinations with just one exception each. In general, results for invariance in loadings were quite good for all four remaining scales, with all but two items out of the total reflecting invariance in loadings for about 95% of countries or higher. Considering that invariance in the loadings (metric invariance) is thought to be a minimum standard for partial invariance, this pattern of favorable results provides basic support for use of the scales in the multiple countries surveyed.

Results for intercept invariance were not quite as positive but were sufficient to provide evidence of comparability of scores. Since invariance in intercepts (scalar invariance) is a more stringent standard than invariance in loadings, more noninvariance is expected. Among the 18 items comprising scales 1, 3, and 4, only two items demonstrated noninvariance in less than 70% of countries (items

25f and 24). The poorest results for intercept noninvariance were for items in scale 2 but were not considered poor enough to restrict interpretation of scale scores overall. More details are provided next.

For country-by-country evaluation, interpretation of alignment results was compared to a standard of the country demonstrating approximate invariance for at least half of the items on the scale. This standard was evaluated separately for the loadings and intercepts. There were a total of 230 invariance tests (29 countries for three scales and 28 countries for one scale, repeated for loadings and intercepts); the standard was met in more than 97% of these tests.

For scales 1 and 3, all countries met the standard for loadings as well as intercepts, showing broad applicability of these scales across the countries in the study. For loadings, there were only two instances where the standard was not met: Slovakia for scale 2 and Lebanon for scale 4. For intercept invariance, the standard was met in all cases except for Italy with scale 2 and three countries with scale 4. These three countries were Croatia, Denmark, and Italy. It is possible that further refinement of the models could have improved alignment results for scale 2, since that scale showed the most noninvariance, although in general results were quite favorable.

Attitudes of Psychologists

Country means and standard deviations for the four remaining scales are provided in Table 6. Table 7 provides effect sizes of each country's difference from the overall mean for each scale for scales 1 through 4. From this point forward, we compare countries to the overall mean rather than directly comparing countries to each other. Instead of simply ranking the country means, significant differences in country means were investigated in the final piece of the Results section. Multilevel models were used to determine whether these ratings differed systematically according to respondent characteristics as well as between countries. We caution readers against explicitly ranking the countries because we have not tested whether individual countries are significantly higher or lower than others. Individual country representatives have been provided with more detailed data to make further evaluations as relevant in their specific context.

For nearly all tests, the effects of the predictors of gender, age, and specialization were significant, with the direction of effects differing across the scales. The coefficients and *p*-values for these analyses are shown in Table 8. For each scale, model-predicted means for various combinations of gender, specialization, and age are provided in Table 9. The starting point for these calculations were as predicted for respondents with an age of 40 years (which is close to the mean of the sample, $m = 41.39$). Significant differences for individual predictors are shown in the table of coefficients but, for example, there was no significance test between female work psychologists and male educational psychologists. Apparent differences in Table 9 should be interpreted accordingly.

TABLE 6
Scale Means and Standard Deviations per Country

Country	Scale 1		Scale 2		Scale 3		Scale 4	
	m	σ	*m*	σ	*m*	σ	*m*	σ
Austria	3.05	0.82	4.09	0.52	2.91	0.56	4.35	0.62
Belgium	2.42	0.74	3.82	0.54	2.83	0.57	4.06	0.64
Brazil	3.19	1.22	4.43	0.49	2.94	0.87	**4.66**	0.46
Bulgaria	3.33	1.01	3.73	0.56	3.08	0.66	4.11	0.56
China	3.54	0.81	3.89	0.52	**3.37**	0.57	4.12	0.50
Croatia	3.24	0.76	3.98	0.46	2.79	0.57	4.34	0.53
Czech Republic	3.50	0.54	3.92	0.54	2.64	0.69	4.18	0.62
Denmark	2.81	0.76	3.82	0.52	2.91	0.60	4.00	0.75
Germany	3.02	0.93	3.81	0.67	3.05	0.77	4.33	0.56
Greece	3.45	0.82	4.41	0.35	*2.06*	0.51	4.33	0.53
Hungary	2.72	0.79	3.76	0.50	2.95	0.67	4.33	0.52
Indonesia	*2.26*	0.58	3.68	0.65	2.80	0.77	4.01	0.77
Israel	3.14	0.69	4.01	0.48	2.70	0.68	4.12	0.57
Latvia	3.20	0.85	3.99	0.57	2.91	0.75	4.21	0.42
Lebanon	**4.16**	0.82	**4.48**	0.42	2.58	0.65	4.29	0.40
Lithuania	2.98	0.88	3.77	0.51	2.80	0.62	4.09	0.53
Netherlands	2.29	0.76	3.61	0.60	2.88	0.61	4.18	0.64
New Zealand	3.12	0.85	3.80	0.48	2.88	0.59	4.32	0.56
Nigeria	3.38	0.90	3.61	0.77	3.02	0.61	*3.97*	0.82
Norway	3.11	0.82	4.02	0.49	2.92	0.61	4.26	0.61
Poland	2.90	0.86	4.14	0.49	2.42	0.63	4.47	0.45
Romania	3.57	0.70	3.95	0.53	2.82	0.68	4.19	0.54
Slovakia	3.16	0.92	*3.49*	0.67	2.24	0.54	4.11	0.56
Slovenia	2.61	0.85	4.06	0.48	2.74	0.58	4.25	0.46
Spain	3.12	0.95	4.00	0.63	2.78	0.71	4.07	0.77
Sweden	2.88	0.90	4.19	0.49	2.74	0.68	4.40	0.66
Turkey	3.64	0.84	4.11	0.56	2.62	0.66	4.11	0.58
United Kingdom	3.34	0.82	3.71	0.61	2.78	0.65	4.12	0.77
Total	3.16	0.90	3.97	0.58	2.76	0.67	4.09	0.70

Note. Boldface = highest scoring country; italics = lowest scoring country.

In summarizing the results one could state that male psychologists had significantly higher ratings than female psychologists on the scales Concerns over incorrect test use and Internet testing, while having significantly lower ratings on Regulations on tests and testing. Compared to younger psychologists, older psychologists had significantly higher ratings on Regulations on tests and testing but significantly lower ratings for the other three scales. Compared to clinical psychologists, educational psychologists had significantly higher ratings on Regulation on tests and testing, and Appreciation of tests, while having significantly lower ratings on Concerns over incorrect test use. Compared to clinical psychologists, work psychologists had significantly higher ratings on Concerns over

TABLE 7
Effect Sizes (*d*) for Country Difference from Grand Mean for Scale Total Scores

Country	Scale 1 *d*	Scale 2 *d*	Scale 3 *d*	Scale 4 *d*
Austria	0.13	0.23	0.27	0.42
Belgium	1.00	−0.28	0.12	−0.05
Brazil	−0.02	0.94	0.21	1.24
Bulgaria	0.17	−0.43	0.48	0.04
China	0.47	−0.15	1.07	0.06
Croatia	0.11	0.02	0.05	0.47
Czech Republic	−0.63	−0.09	−0.17	0.15
Denmark	−0.46	−0.29	0.25	−0.12
Germany	0.15	−0.24	0.38	0.43
Greece	−0.35	1.26	−1.37	0.45
Hungary	−0.56	−0.42	0.28	0.46
Indonesia	1.55	−0.45	0.05	−0.10
Israel	−0.03	0.08	−0.09	0.05
Latvia	0.05	0.04	0.20	0.29
Lebanon	−1.22	1.21	−0.28	0.50
Lithuania	0.20	−0.39	0.06	−0.00
Netherlands	−1.14	−0.60	0.20	0.14
New Zealand	−0.05	−0.35	0.20	0.41
Nigeria	0.24	−0.47	0.43	−0.15
Norway	0.06	0.10	0.26	0.28
Poland	0.30	0.35	−0.54	0.84
Romania	0.59	−0.04	0.09	0.19
Slovakia	0.00	−0.72	−0.96	0.04
Slovenia	0.65	0.19	−0.03	0.35
Spain	0.04	0.05	0.03	−0.03
Sweden	0.31	0.45	−0.03	0.47
Turkey	0.57	0.25	−0.21	0.03
United Kingdom	−0.22	−0.43	0.03	0.04

incorrect test use and Internet testing, while having significantly lower ratings on Regulations on tests and testing.

Comparisons Between Countries

In addition to differences in scale means for respondents with different characteristics, it is also desirable to determine which individual countries have statistically different mean scores for the scales. Multilevel models were estimated for each scale, including countries as random effects. Significance tests were conducted by outputting the estimate for each country's random intercept, along with confidence intervals for these estimates. The confidence intervals associated with the countries' random effects were then used to determine which countries had means significantly higher

TABLE 8
Multilevel Model Results: Fixed Effects Coefficients

	Scale 1 Concerns over Incorrect Test Use	Scale 2 Regulations on Tests and Testing	Scale 3 Internet Testing	Scale 4 Appreciation of Tests
Intercept	3.11**	3.95**	2.72**	4.18**
Male	0.10**	−0.10**	0.12**	−0.02*
Education	−0.12**	0.04**	−0.02	0.09**
Work	0.17**	−0.01	0.26**	0.01
Age	−0.04**	0.04**	−0.08	−0.02**

or significantly lower than the overall mean. That is, each country was grouped into one of three categories: significantly greater than the mean, significantly lower than the mean, or not significantly different from the mean. This system creates a parsimonious interpretation that is more straightforward than pairwise difference tests between individual countries in some traditional analyses.

As previously described, there were major differences between countries in the demographic composition of the sample and of the population. Therefore, some differences between country means on the scales might be attributable to patterns of demographic variables rather than unique characteristics of the countries themselves. The question of whether countries differ overall on the scale means and whether the countries differ after taking into account background characteristics are both relevant but are conceptually distinct.

For the country comparisons, the models were estimated two ways: both with and without inclusion of the demographic characteristics. Models that included the demographic characteristics are labeled "conditional" in the current study, and models that do not include the demographic characteristics are labeled "unconditional." As with

TABLE 9
Model-Predicted Scale Means by Demographic Variables

	Scale 1	Scale 2	Scale 3	Scale 4
Female, Clinical	3.11	3.95	2.72	4.18
Female, Education	2.99	3.99	2.70	4.27
Female, Work	3.28	3.94	2.99	4.19
Male, Clinical	3.21	3.85	2.84	4.16
Male, Education	3.09	3.89	2.82	4.25
Male, Work	3.38	3.84	3.11	4.17
Female, Age 50	3.07	3.99	2.64	4.15
Female, Age 30	3.15	3.91	2.81	4.20
Male, Age 50	3.17	3.89	2.76	4.14
Male, Age 30	3.25	3.82	2.93	4.18

Note. Standard comparison is for Clinical, Age 40. Other combinations can be calculated from the values in Table 8.

the intercept reported in the previous section, the mean scores for the conditional models represent the mean within the country for a 40-year-old female clinical psychologist. Recall that these values were chosen because they were near the mean age and were the most common gender and specialization. The results tables for this section include both the conditional and unconditional models. The graphs report results for only the conditional models to minimize apparent differences associated with demographic characteristics rather than unique country features.

Table 10 shows the results of the significance testing comparing countries. For each of the four reported scales, several countries were significantly higher or significantly lower than the overall mean. For the most part, there was substantial overlap between the results from the conditional and unconditional models. In numerous instances, countries that showed significant differences in the unconditional model were not significantly different from the mean in the conditional model. In such cases, apparent mean differences are likely related to differences in sample demographics rather than unique country features. In a small number of instances, countries that

TABLE 10
Countries Significantly Different Than Overall Intercept in Multilevel Models

	Scale 1	Scale 2	Scale 3	Scale 4
High	Bulgaria*	Austria*	Austria*	Austria*
	China	Brazil	Bulgaria	Brazil*
	Czech Republic	Greece	China	Croatia
	Greece	Italy	Denmark	Germany
	Italy	Lebanon	Germany	Poland
	Lebanon	Norway	Hungary	Sweden
	Nigeria*	Poland	Netherlands	*Norway*
	Romania	Slovakia*	Nigeria*	
	Turkey	Sweden	Norway	
	United Kingdom	Turkey		
		Spain		
Low	Belgium	Belgium	Czech Republic	Belgium
	Denmark	Bulgaria	Greece	Denmark
	Hungary	Denmark	Italy	Italy
	Indonesia	Germany*	Poland	Nigeria*
	Netherlands	Hungary	Slovakia	Spain
	Poland	Indonesia	Turkey	United Kingdom*
	Slovenia	Lituania		
	Sweden	Netherlands*		
		Nigeria		
		United Kingdom		
		Slovakia		

Notes. Lack of special formatting indicates country difference was significant for conditional and unconditional models. Italics indicate country difference was significant only for conditional model. Asterisks indicate country difference was significant only for unconditional model.

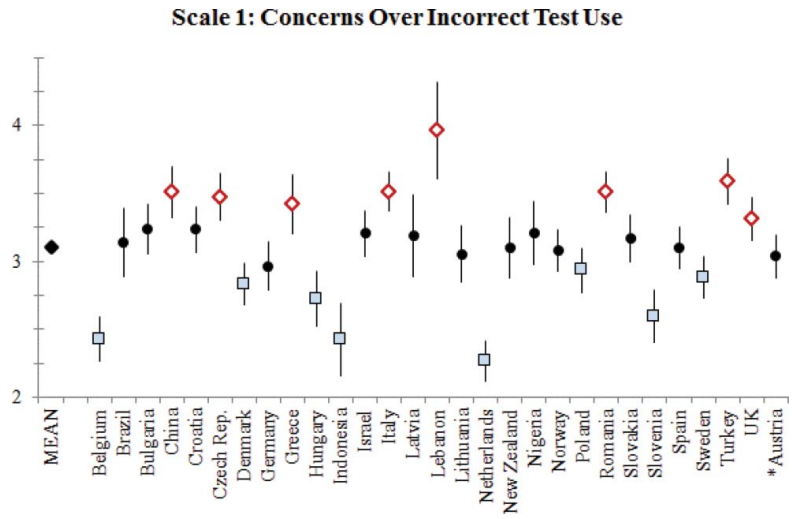

FIGURE 2
Scale One: Random intercepts and confidence intervals for countries in conditional model.

were not significantly different from the mean in the unconditional model were significantly different for the conditional model.

Figures 2–5 show the individual random intercept estimate and confidence intervals for each country. Smaller confidence intervals indicate more precision in the

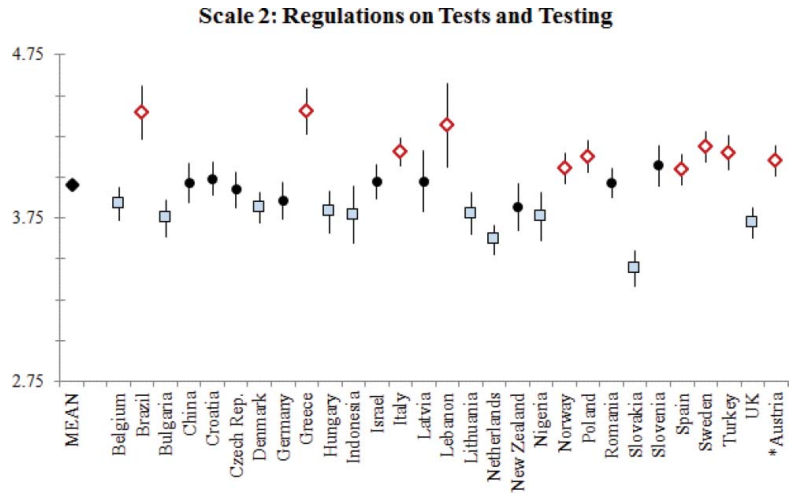

FIGURE 3
Scale Two: Random intercepts and confidence intervals for countries in conditional model.

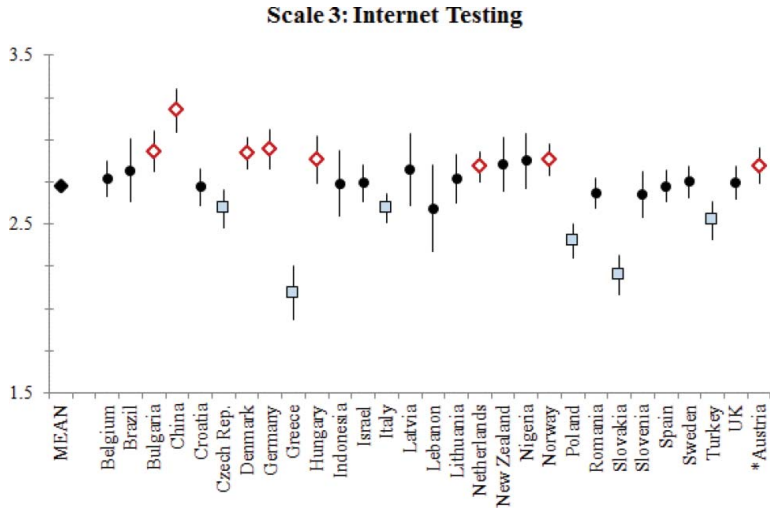

FIGURE 4
Scale Three: Random intercepts and confidence intervals for countries in conditional model.

estimate for that country, due to larger sample or smaller variance. In the figures, diamond markers indicate country random intercepts significantly higher than the overall intercept, and square markers indicate significantly lower random intercepts. For each scale, many countries were not significantly different from the mean.

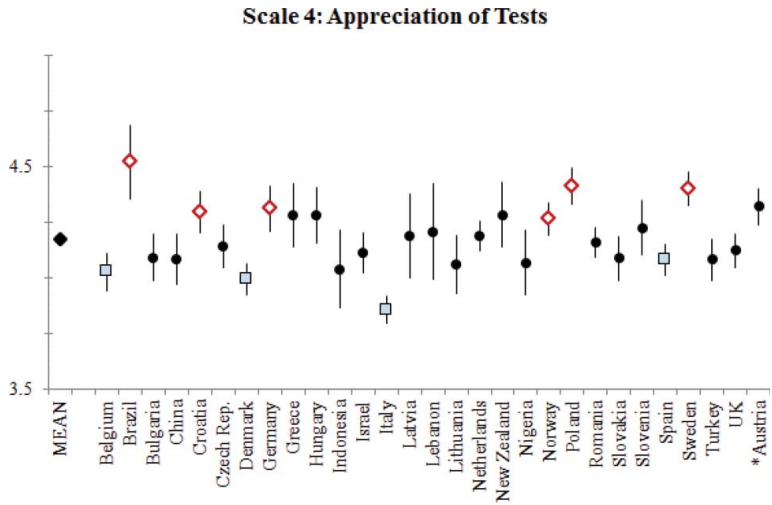

FIGURE 5
Scale Four: Random intercepts and confidence intervals for countries in conditional model.

TABLE 11
Items with *d*-values > 1.6 Between Extreme Scoring Countries

Scale	Item Text	Lowest Scoring		Highest Scoring		Lowest vs Highest	
		Country	Mean	Country	Mean	Difference	*d*
1	Not restricting test administration to qualified personnel	Belgium	1.59	Lebanon	4.21	2.62	1.9
	Making interpretations which go beyond the limits of the test	Indonesia	1.92	Lebanon	4.16	2.24	1.8
	Not considering errors of measurement of a test score	Indonesia	2.15	Lebanon	4.11	1.96	1.7
2	Anyone who can demonstrate their competence as a test user (whether a psychologist or not) should be allowed to use tests	China	3.35[1]	Italy	1.34[1]	1.81	1.7
	Our National Psychological Association should take a more active role in the regulation and improvement of test use	Slovakia	3.07	Turkey	4.71	1.64	1.7
	The use of psychological tests should be restricted to qualified psychologists	Indonesia	3.29	Lebanon	4.89	1.60	1.6
3	Test administration over the Internet has many advantages compared with paper-and-pencil administration	Greece	1.35	China	3.90	2.51	2.2
	If properly managed, the Internet can greatly improve the quality of test administration	Greece	1.97	China	4.00	2.03	1.9
	Computer-generated interpretive reports do not have any validity	Slovakia	4.15[1]	Germany	2.38[1]	1.72	1.6
5	The training received in psychology bachelors' degree courses is sufficient for the correct use of most tests	Greece	1.11	Indonesia	3.23	2.12	2.0
	My current knowledge with regard to tests is basically that which I learned on my psychology degree course	Greece	1.39	Lithuania	3.77	2.38	1.9
	The training received in psychology masters' degree courses is sufficient for the correct use of most tests	Greece	2.09	Germany	4.08	1.99	1.7

[1]For reverse items the highest scoring country is mentioned in the column "lowest" and vice versa.

Interpretation of the results and differences between countries on item level may be worthwhile for national associations and others to guide their actions. However, a table of the means and standard deviations on 32 items of 29 countries would take too much space.[3] In addition, the results on some items might be of more interest than those on others. Therefore, for each scale only the results of the extreme scoring countries on the items with *d*-values greater than 1.6 between these countries are given

[3]A table with means and standard deviations on all items of all countries can be obtained from the first author.

(see Table 11, in which also the items of scale 5 are included). As item scores show more variability than scale scores, twice the value indicating a strong effect ($d = 0.8$, Cohen, 1988) was chosen as the cut-off to include items in the table.

The *d*-values in Table 11 show very large difference between extreme-scoring countries for these selected items. (Note that these results are for individual items, rather than the full scales.) Particularly of interest are the very high score of Lebanon on the items in scale 1 (compared to the quite low scores in Belgium and Indonesia); the very high scores (near the top of the scale) of Italy, Turkey and Lebanon on the items in scale 2; and the very low scores (near the bottom of the scale) of Greece on the items in scale 3 and 5. It is also of note that Italy, Greece, Turkey, and Lebanon, which are geographic neighbors, were the most extreme scoring for these items on three of the five scales. The differences on the items in scale 4 were much smaller, with *d*-values near one between the extreme scoring countries.

DISCUSSION

An important observation obtained from the data is the positive attitude of the participating psychologists toward the use of tests. Although the multilevel analysis shows that some countries significantly differ from the mean, the differences are small. The conditional model shows that for scale 4 the mean in all countries but one (Italy) is above 4 on the 1 to 5 scale (see Figure 5). Therefore, it can be concluded that this positive attitude applies to all 29 countries. However, it should be noted that the means may be inflated by an overrepresentation of test users (compared to nontest-using psychologists) in the sample. It seems plausible to expect that test users are more willing to participate in a survey on test attitudes and might show a more positive attitude than nontest users. Nonetheless, considering the very high values and the stable pattern across the 29 countries with varying response ratios, the conclusion stated in Muñiz and colleagues (2001, p. 208) can be repeated: "Psychologists have no hesitation in using tests in the exercise of their profession . . . considering them as a helpful tool."

In contrast to the positive overall views of tests, most countries had means just below the midpoint of 3 on scale 3, addressing attitudes about Internet tests, with relatively small differences between most countries. Only one country (China) scores clearly above this midpoint (see Figure 4). The results show a widespread lack of enthusiasm for Internet tests, expressing only moderate appreciation for such tests. Considering the age distribution of the sample compared to the timeline of Internet testing, it is likely that Internet testing was not a major part of the curriculum for a majority of psychologists in the sample.

The differences between the country means on scale 1, Concern over incorrect test use, and scale 2, Regulations on tests and testing, are rather large, with

differences between the extreme scoring countries of about 1.5 *d*. Countries scoring high on concerns over incorrect test use also argue for stricter regulations on tests and testing (such as Lebanon), and vice versa (such as Indonesia and the Netherlands, see Figures 2 and 3); there was a correlation of 0.42 between the country means on these scales. This suggests a substantial concern of test users when the correct use of tests is at risk and a more relaxed attitude when test users think that tests are used in a proper manner in general.

Overall, gender differences and differences between fields of specialization are much smaller than those between countries. On the scales Regulations and Appreciation, the effects of these characteristics are small, with differences between extreme scoring groups of about 0.2 *d*. On the scales addressing concerns over incorrect test use and Internet testing, there are medium size effects of about 0.5 *d*. On these scales, work psychologists show both greater concern over the incorrect use of tests as well as a greater appreciation of Internet testing, compared to those of educational and clinical psychologists. An explanation for the first effect may be that in the work field, relatively more nonpsychologists use tests; an explanation for the second effect may be that Internet testing is already more common in the work field than educational and clinical testing, which may lead to greater acceptance (see also Hambleton, 2004).

We examined how the isolated incidences of noninvariance in the alignment results might affect interpretation of the overall study results and comparisons. Since invariance for the loadings is often thought to be a minimum standard for partial invariance, we were especially cautious about interpretation of the results in cases where this standard was not met. For Lebanon in scale 4 (appreciation of tests), the country's mean for the scale was not significantly higher or lower than the mean in the earlier analysis and thus conclusions were likely to be muted in any case. For Slovakia in scale 2 (regulations on tests and testing), Slovakia's scale mean was markedly below the other countries. The alignment results for this scale suggest that the seemingly low score may be influenced by noninvariance of the loadings. Similarly, Italy had the lowest observed scale mean for scale 4, which may have been influenced by noninvariance of the intercepts.

Recall that for scale 2, Croatia, Denmark, and Italy failed to meet the standard for intercept invariance. Croatia and Denmark had observed means for this scale quite near the overall mean, while Italy's observed mean was somewhat higher than the overall mean. Representatives for these countries would be cautioned against making broad comparative conclusions in these cases. Overall the multicountry trends remain salient for this scale as well as the other three scales examined.

Implications for Practice

Despite the general positive attitude toward tests, the negative attitude with respect to Internet testing and the concerns of psychologists regarding the

incorrect use of tests is concerning. These results urge action of national associations, international bodies such as ITC and EFPA, faculties of psychology, test authors, and test publishers. Because it seems inevitable tests will be administered by computers more frequently (see also Hambleton, 2004), authors, publishers, and relevant organizations should invest time and energy in dispelling the distrust in Internet testing as well as ensuring appropriate use of such tests. These efforts should focus in particular on psychologists in the educational and clinical field. Important in this respect is also the publication and dissemination of the *ITC Guidelines on Computer-Based and Internet-delivered Testing* (ITC, 2005), as this publication may meet the need for more information on this issue. It is also the duty of the national associations to convince the faculties of psychology to bring their curricula up to date.

A potential method for decreasing concern about incorrect use of testing could be the adoption of new regulations by national associations. In this respect, the publication of a set of standards for the delivery of services for the assessment of people in work and organizational settings by the International Organization of Standardisation (2011a, 2011b), the ITC *International Guidelines on Quality Control in Scoring, Test Analysis, and Reporting of Test Scores* (2012) and the EFPA *Standards for Test Use* (2012) and *Standards for Psychological Assessment* (2013) could guide national associations in their efforts. However, a study by Rios and Sireci (2014) showed that the simple publication of guidelines and regulations is not enough to change behavior in practice. Much effort should be invested in promoting the implementation of guidelines and regulations, for example, by ensuring that these regulations get a prominent place in the curriculum of psychology. Stakeholders in each country could also use the detailed results of this survey to direct their actions.

Strengths, Limitations, and Future Research

The strong element of this study is the large number of countries that participated. It is the first systematic study that assesses attitudes of psychologists worldwide. However, a weak point is that the number of participating countries outside Europe is limited. Moreover, although many more than 29 countries were invited to participate, there was a type of self-selection at work depending on the willingness of a country to participate. Therefore, it is not appropriate to generalize the results of this study to other countries (or to psychologists all over the world). In addition, the sample sizes of some countries are rather small, which may limit the trustworthiness of the conclusions for the countries concerned and for the differences of these countries (e.g., Indonesia, Latvia and Lebanon having N's < 50) from the overall mean. In addition, the power to detect mean differences from other countries may be low for countries with small sample sizes.

A further limitation concerns the response rate and consequently the representativeness of the sample. As was shown in Table 1, the response rate of the total sample is 11.3%, ranging from a low 3.4% in Germany to a high 42.2% in Slovakia. A complicating issue is that there was substantial variation in the way the survey was administrated and distributed in participating countries, ranging from digital versions sent to all members of the national psychological association to paper versions handed out at local conferences. Although it showed that mode of administration per se had no effect on the results, the low general response rate, combined with the variation in way of distribution and approaching respondents over countries, inhibits generalizing the conclusions to all psychologists in the 29 countries.

Although the check for representativeness on two of the inventoried background variables (gender and professional field) showed significant differences, the results seem to be rather reassuring, because differences are small, as shown by the effect sizes. Another reassuring aspect was the substantial overlap between results from the conditional and unconditional models. All in all, one should be careful in generalizing the results to groups that exceed the sampling frame (i.e., members of the psychological association, or psychologists visiting conferences) used in a particular country.

Hambleton (2004, 2006) mentioned six overarching areas that will attract the attention of researchers and professionals in the coming years. These areas concern the internationalization of testing, the use of new psychometric models and technologies to generate and analyse tests, the appearance of new item formats derived from computer and multimedia advances, the further development of computerized tests and testing by the Internet, the development of systems used to report the results to users or others who may have a need or right to see them, and the growing demand for training by diverse professionals (not just psychologists) who use assessment. In light of these continued changes, it will be challenging for universities who educate psychologists, ITC/EFPA, national psychological associations, and psychologists themselves to keep their curricula, regulations, and knowledge up-to-date.

The opinions of psychologists can play an important role in this continuing process. Therefore, repeating this survey after some years seems advisable (e.g., an interval of 10 to 15 years would likely be long enough to detect changes). In the preparation of a future survey much effort should be invested in having more countries outside Europe participate (industrialized as well as developing countries), getting a higher response rate (which may be increased by using a more intense and personal approach), and conducting a more uniform method of administration (which may be facilitated by the growing dissemination of the Internet in less developed countries). The more these conditions are fulfilled, the more confidence stakeholders in the field of testing can draw on these results to inform their actions (i.e., by the provision of directed training and information).

ACKNOWLEDGMENTS

We thank the members of the EFPA Board of Assessment for their help in the different phases of this work, the "Friends of the ITC" who participated in the second phase of this study, and the National Psychological Associations that facilitated the administration of the questionnaire. Our special thanks go to Dragos Iliescu who contacted the Friends of the ITC.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 495–508.

Bartram, D. (2011). Contributions of the EFPA Standing Committee on Tests and Testing (SCTT) to standards and good practice. *European Psychologist*, *16*, 149–159.

Bartram, D., & Coyne, I. (1998). Variations in national patterns of testing and test use: The ITC/EFPPA international survey. *European Journal of Psychological Assessment*, *14*, 249–260.

Bartram, D., & Hambleton, R. K. (Eds.) (2006). *Computer-based testing and the Internet*. Chichester, UK: Wiley and Sons.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

European Federation of Psychologists' Associations. (2012). *Standards for Test Use: Work, Education, and Health & Social Care, Levels 1, 2 and 3*. Brussels, Belgium: Author.

European Federation of Psychologists' Associations. (2013). *Standards for Psychological Assessment: 2013 Work, Education, and Health & Social Care, Levels 1, 2 and 3. Performance Requirements, Context definitions and Knowledge & Skill specifications for the three EFPA levels of qualifications in psychological assessment*. Brussels, Belgium: Author.

Evers, A. (2012). The internationalization of test reviewing: Trends, differences, and results. *International Journal of Testing*, *12*, 136–156.

Evers, A., Muñiz, J., Bartram, D., Boben, D., Egeland, J., Fernández-Hermida, J. R., et al. (2012). Testing Practices in the 21st Century: Developments and European Psychologists' Opinions. *European Psychologist*, *17*, 300–319.

Evers, A., Zaal, J. N., & Evers, A. K. (2002). Ontwikkelingen in het testgebruik van Nederlandse psychologen [Developments in test use of Dutch psychologists]. *De Psycholoog*, *37*, 54–61.

Eyde, L. D., Moreland, K. L., Robertson, G. J., Primoff, E. S., & Most, R. B. (1988). Test user qualifications: A data-based approach to promoting good test use. In *Issues in scientific psychology*. Washington, DC: American Psychological Association.

Eyde, L. D., Robertson, G. J., Krug, S. E., Moreland, K. L., Robertson, A. G., & Shewan, C. M., et al. (1993). *Responsible test use. Case studies for assessing human behavior*. Washington, DC: American Psychological Association.

Fine, S. (2013). A critical look at psychological testing in Israel and comparisons with its European neighbors. *International Journal of Testing*, *13*, 249–271.

Hambleton, R. K. (2004). Theory, methods, and practices in testing for the 21st century. *Psicothema*, *16*, 696–701.

Hambleton, R. K. (2006, March). *Testing practices in the 21st century*. Key Note Address, University of Oviedo, Spain.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.) (2005). *Adapting educational and psychological tests for cross-cultural assessment*. London: Lawrence Erlbaum.

International Organization for Standardization. (2011a). *Assessment service delivery—Procedures and methods to assess people in work and organizational settings—Part 1: Requirements for the client (ISO 10667-1:2011, IDT)*. Geneva, Switzerland: Author.

International Organization for Standardization. (2011b). *Assessment service delivery—Procedures and methods to assess people in work and organizational settings—Part 1: Requirements for service providers (ISO 10667-2:2011, IDT)*. Geneva, Switzerland: Author.

International Test Commission. (2001). International guidelines on test use. *International Journal of Testing*, *1*, 95–114.

International Test Commission. (2006). International guidelines on computer-based and Internet-delivered testing. *International Journal of Testing*, *6*, 143–172.

International Test Commission. (2012). *International guidelines on quality control in scoring, test analysis, and reporting of test scores*. Retrieved from www.intestcom.org

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: Taylor and Francis Group.

Muñiz, J., & Bartram, D. (2007). Improving international tests and testing. *European Psychologist*, *12*, 206–219.

Muñiz, J., Bartram, D., Evers, A., Boben, D., Matesic, K., Glabeke, K., Fernández-Hermida, J.R., & Zaal, J. (2001). Testing practices in European countries. *European Journal of Psychological Assessment*, *17*, 201–211.

Muñiz, J., Prieto, G., Almeida, L., & Bartram, D. (1999). Test use in Spain, Portugal and Latin American countries. *European Journal of Psychological Assessment*, *15*, 151–157.

Muthén, B., & Asparouhov, T. (2013). New methods for the study of measurement invariance with many groups (Mplus Technical Report). Retrieved from http://www.statmodel.com

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Rios, J. A., & Sireci, S. G. (2014). Guidelines versus practices in cross-lingual assessment: A disconcerting disconnect. *International Journal of Testing*, *14*, 289–312.

SAS Institute Inc. (2011). *SAS/STAT 9.3 User's Guide*. Cary, NC: SAS Institute.

APPENDIX

EFPA Questionnaire on Test Attitudes of Psychologists – EQTAP

On the initiative of the *International Test Commission* (*ITC)* and the *European Federation of Psychologists Associations* (*EFPA) Standing Committee on Tests and Testing*, a survey is being carried out in countries all around the world in order to find out the opinion of psychologists on various aspects related to tests and testing. This information will be of assistance in making decisions to improve test use in our country. We would greatly appreciate your devoting a few minutes to giving us your sincere opinion about a series of matters related to tests. Your responses should relate to your understanding of the situation in your area of professional specialisation. Thank you very much for your cooperation. The survey is anonymous.

General data

Age: …….. year (indicate)

Sex: male / female (circle the right answer)

Professional speciality: Clinical/Health Education Work Other .....(indicate)

Questionnaire

Your responses are to be made on a scale of 1-5:
if you *totally disagree* with the statement, circle 1; if you *totally agree* with the statement, circle 5. Use the numbers 2, 3 and 4 for intermediate opinions.

1	The training received in psychology bachelors' degree courses is sufficient for the correct use of most tests	1	2	3	4	5
2	The training received in psychology masters' degree courses is sufficient for the correct use of most tests	1	2	3	4	5
3	The *ITC*, the *EFPA*, or any other international organization should establish a global system to accredit the certification of test users	1	2	3	4	5
4	Professionals are provided with sufficient information (independent reviews, research, documentation, etc.) on the quality of tests published in my country	1	2	3	4	5
5	In my professional field computer-based testing is progressively replacing paper and pencil tests	1	2	3	4	5
6	My current knowledge with regard to tests is basically that which I learned on my psychology degree course	1	2	3	4	5
7	Test administration over the Internet has many advantages compared with paper-and-pencil administration.	1	2	3	4	5
8	The use of psychological tests should be restricted to qualified psychologists	1	2	3	4	5
9	While non-psychologists may administer and score tests, interpretation and feedback should be restricted to psychologists	1	2	3	4	5
10	Computer-generated interpretive reports do not have any validity	1	2	3	4	5
11	Standards [e.g., those of the EFPA, or of the American Psychological Association (APA)] defining the minimum technical qualities of a test should be enforceable	1	2	3	4	5

12	Legislation is needed to control the more serious abuses of testing	1	2	3	4	5
13	Test administration over the Internet sets some test takers at a disadvantage	1	2	3	4	5
14	Anyone who can demonstrate their competence as a test user (whether a psychologist or not) should be allowed to use tests	1	2	3	4	5
15	If properly managed, the Internet can greatly improve the quality of test administration	1	2	3	4	5
16	Controls on tests and testing should be minimal, as controls discourage the development of new ideas and new procedures	1	2	3	4	5
17	The privacy of the test taker is not protected when testing by Internet	1	2	3	4	5
18	Publishers should be allowed to sell whatever tests they think fit	1	2	3	4	5
19	Our National Psychological Association should take a more active role in the regulation and improvement of test use	1	2	3	4	5
20	Testing over the Internet opens the way to fraud	1	2	3	4	5
21	I use tests regularly in the exercise of my profession	1	2	3	4	5
22	Tests constitute an excellent source of information if they are combined and complemented with other psychological data	1	2	3	4	5
23	Used correctly, tests are of great help to the psychologist	1	2	3	4	5
24	All things considered, in the last decade tests and testing practices have improved in my country	1	2	3	4	5

Your responses on the next 8 questions are to be made on a scale of 1-5:

1= very rarely; 5 = very frequently. Use the numbers 2, 3 and 4 for intermediate opinions.

25	Indicate the frequency with which you believe the following test-use problems occur within your professional speciality in your country:					
a	Making photocopies of *copyrighted* materials	1	2	3	4	5
b	Making evaluations using inappropriate tests	1	2	3	4	5
c	Not keeping up with the field	1	2	3	4	5
d	Failing to check one's own interpretations with others	1	2	3	4	5
e	Not considering errors of measurement of a test score	1	2	3	4	5
f	Not restricting test administration to qualified personnel	1	2	3	4	5
g	Not taking into account conditions that cast doubt on reported validity for a local situation	1	2	3	4	5
h	Making interpretations which go beyond the limits of the test	1	2	3	4	5

26. Name the three tests you use most frequently in the exercise of your profession (if you never use tests, you can skip this question):

1 .....
2 .....
3 . .....

Comments. Please make any additional comments you think appropriate (you may include extra sheets, if necessary):

.....
.....
.....

Thank you very much for your cooperation.