Clínica y Salud

Is there a social desirability scale in the MMPI-2-RF?

Fernando Jiménez-Gómez[a]*, Guadalupe Sánchez-Crespo[a], and Amada Ampudia-Rueda[b]

[a]University of Salamanca, Spain
[b]National Autonomous University of Mexico, Mexico

ABSTRACT

The purpose of this study is to search for a validity scale for detecting social desirability bias in the MMPI-2-RF. To that end, data from scales considered as underreporting on the MMMPI-2-RF, such as the Edwards and Wiggins Social Desirability (Esd and Wsd respectively) scales, the Other Deception (ODecp) scale of Nichols & Greene, the Superlative (S) scale of Butcher and Han, and Uncommon Virtues (L-r) and Adjustment Validity (K-r) scales, was analyzed comparatively. The sample was taken from the Spanish adaptation of the MMPI-2 database, with the corresponding item selection made from the restructured MMPI-2-RF. Two groups of participants were established: The honest group and the dissimulator group, participants of the latter being instructed to give socially desirable responses. A Receiver Operating Characteristic (ROC) methodology was used to suggest a scale that offers more diagnostic accuracy.

© 2013 Colegio Oficial de Psicólogos de Madrid. All rights reserved.

¿Existe una escala de deseabilidad social en el MMPI-2-RF?

RESUMEN

El objetivo del presente estudio es buscar una escala de validez que detecte la deseabilidad social en el MMPI-2-RF. Para ello se analizaron comparativamente los datos ofrecidos por las escalas de Deseabilidad Social de Edwards (Esd), de Wiggins (Wsd), la escala de Engaño de Nichols & Greene (ODecp), la Superlativa de Butcher y Han (S) y las de Virtudes Inusuales (L-r) y Validez del Ajuste (K-r), propuestas como minimización de síntomas en el MMMPI-2-RF. La muestra fue obtenida de la base de datos del MMPI-2 para la adaptación castellana con la correspondiente selección de los ítems realizada con la forma reestructurada del MMPI-2-RF. Se establecieron dos grupos de participantes: sinceros y disimuladores, siendo éstos últimos instruidos para contestar al cuestionario de forma socialmente deseable. Se utilizó la metodología Receiver Operating Characteristic (ROC) para proponer la escala que presentara mejor precisión diagnóstica.

© 2013 Colegio Oficial de Psicólogos de Madrid. Todos los derechos reservados.

Simulation is a deliberate behavior, the purpose of which is deceiving or lying about a given event in order to obtain economic or psychological gain or avoid duty or responsibility. The DSM-IV-TR defines it as an "intentional fabrication of false or grossly exaggerated physical or psychological symptoms, motivated by external incentives" (American Psychiatric Association, 2000).

With certain motivation, some people's interest *to look good* to others arises in specific circumstances or settings. So they exhibit basically two types of behavior with the sole purpose of obtaining from others a positive self-image (gain or benefit): They either credit themselves with socially desirable behavior or deny undesirable behavior. In this sense, simulation relates to *dissimulation*, the often unintentional and not always conscious behavior of conveying to others a different positive image from the actual one.

Individuals' interest in presenting themselves as *socially desirable* is not new. Since the beginnings of this concept in the 30's with Bernreuter's work (1993) and over the last 50 years, social desirability has become an alarming, recurring topic of interest to Psychology professionals and behavioral assessors (Andrews & Meyer, 2003). Trying to project a socially desirable self-image to others is an intrinsic characteristic of individual personality itself (Eysenck & Eysenck, 1976). Nevertheless, when it comes to exceeded normality limits we, as assessors, must look closely in order to detect this type of bias.

Existing research on social desirability shows particular incidence depending on its different denominations and definitions. For Bagby and Marshall (2004), self-deception is characterized as a general willingness to think about *oneself* in slightly more favorable terms. "Impression management" is a "deliberate attempt to distort

*Correspondence concerning this article should be sent to Fernando Jiménez Gómez. Facultad de Psicología. Universidad de Salamanca. Avda de la Merced, 101. 37005 Salamanca. E-mail: fjimenez@usal.es

responses in order to make a favorable impression *on others*", as defined by Barrick & Mount (1996, p. 262). Crowne & Marlowne (1960) put it simply as presenting oneself in a favorable way. The reasons and variables are many, both personal and situational, and may be causing socially desirable responding.

Given the importance of ensuring data reliability, some researchers have worked both with social desirability bias detection scales (Edwards, 1962; Elvekrog & Vestre, 1963; Fordyce, 1956; Hanley, 1956; Heilbrun, 1964) and simulator groups in a variety of settings and with different scales (Arce, Fariña, Carballal, & Novo, 2006; Graham, Watts, & Timbrook, 1991; Jiménez & Sánchez, 2003; Rogers, 2008; Rogers & Bender, 2003). Other authors have revisited the analysis of dissimulation detection Specificity and Sensitivity using the ROC method (Receiver Operating Characteristic) in order to detect various biases in provided data (Nicholson, Mouton, Bagby, & Buis, 1997; Pelegrina, Ruiz-Soler, & Wallace, 2000; Sellbom and Bagby, 2010; Wygant et al., 2011).

The Minnesota Multiphasic Personality Inventory-2 (MMPI-2, Butcher, Graham, Tellegen, Dalhstron, & Kaemmer, 2001) has been long used not only in forensics (Archer, Buffington-Vollum, Vauter-Strendy, & Handel, 2006; Borum & Grisso, 1995; Lees-Haley, 1992), but also in mental health and clinical practice (González, Santamaría, & Capilla, 2012).

Oddly enough, due to the important booming of research on this topic, the inventory has passed from being a technique which purportedly could be faked easily due to its self-reporting structure, to becoming the preferred technique due to its accuracy in detecting simulation and bias.

Specifically, the Minnesota Multiphasic Personality Inventory (MMPI), and its newest version (MMPI-2), has increased interest for the development and inclusion of a battery of validity scales detecting psychopathology exaggeration or minimization, often referred to as *faking-bad* and *faking-good* respectively, creating a "second generation" of scales detecting these patterns of response bias (simulation and dissimulation) on the MMPI-2. The validity scale configuration in the newest review of the MMPI-2 (Butcher et al., 2001) encompasses traditional scales (L, F, and K), and experimental scales, such as Edwards Social Desirability (Edwards, 1957), Wiggins Social Desirability (Wiggins, 1959), Other Deception (Nichols & Greene, 1991), and the Superlative Scale of Butcher and Han (Butcher & Han, 1995).

For the development of their scales, Wiggins (1959), Edwards (1957), Butcher and Han (1995), and Nichols and Greene (1991) used the 567 items comprising the MMPI-2, the number of items and items used varying from one scale to the other (33 for Wiggins', 37 for Edwards', 50 for Butcher and Hand's, and 33 for Nichols and Greene's). Wiggins and Edwards scales seem to differ in item configuration, unlike the scale of Nichols and Greene, which shows a high item overlapping (70.6%) with Wiggins scale.

The MMPI-2-RF (Ben-Porath & Tellegen, 2008; Santamaría, 2009) is the restructured form of the MMPI-2 (Ávila & Jiménez, 1999; Butcher et al., 2001). This restructuration has some really innovative features: A smaller item number (338 items taken from the 567 item MMPI-2), shorter administration time (approx. 35-50 minutes), more accuracy of protocol validity evaluation (as it includes the results from research conducted during the last decade on simulation and dissimulation on the MMPI-2), reduction of redundancy and interpretative inconsistency, and adapted to modern conceptions of clinical personality. The MMPI-2-RF interpretative structure has been updated significantly and divided into two big sections, each one featuring relevant subsections: 1) protocol validity, which includes the assessment of response consistency and possible presence of symptom exaggeration (*overreporting*) (Sellbom & Bagby, 2010) and symptom minimization (*underreporting*) patterns, and 2) the 42 substantive scales are grouped to specifically analyze the possible presence of somatic/cognitive, emotional, behavioral, and

thought effects, as well as interpersonal relationships, assessed subject interests, diagnostic and therapeutic considerations (Santamaría, 2009).

MMPI-2-RF validity scales include eight scales, the most part being a review of those in the MMPI-2, and only one of them being completely original, the Infrequent Somatic Responses (Fs). In this review, minor modifications were made to validity scales, such as renaming (e.g., the Lie [L] scale is now denominated "Uncommon Virtues") or adding an "r" (revised) to their acronym to make a distinction (e.g., F-r, L-r, K-r).

The interest for this investigation work on social desirability has derived from Ben-Porath and Tellegen's new restructuration (2008) of the MMPI-2 (MMPI-2-RF) because of three main reasons: 1) responses in this type of questionnaire can be easily altered and manipulated according to the personal interests of respondents; 2) both in the field of simulation/dissimulation in forensic evaluations and legal medicine and mental disorders, detecting the accurate validity and reliability of data provided by administered tests is required; and 3) the MMPI-2-RF, as its predecessor (the MMPI-2), also lacks a specific scale for this variable to help us detect social desirability bias.

The purpose of this study is to make a comparative analysis of three specifically denominated scales of the MMPI-2-RF to do the detecting of social desirability bias: Edwards Social Desirability (Edwards, 1957), Wiggins Social Desirability (Wiggins, 1959), and Other Deception (Nichols & Greene, 1991). Butcher and Han's (1995) Superlative Scale (S) and dissimulation (to project a good self-image, or underreporting) detecting scales as proposed by the MMPI-2-RF, such as Uncommon Virtues (L-r), and Adjustment Validity (K-r), are also included in this analysis. This is expected, on the one hand, to show the tester if there exists a scale detecting social desirability bias (even if it is not designed expressly to do so by name) and, on the other hand, to evaluate the scale showing more diagnostic accuracy through various statistical analyses, including a ROC (Receiver Operating Characteristic) curve analysis with the Area Under the Curve, Sensitivity, Specificity, and Positive and Negative Predictive Power (Preti et al., 2007). In order to achieve this, the existing database of subjects who have completed the MMPI-2 scales was used. Responses from two groups of participants who completed the MMPI-2 under different instructions have been examined: the "honest" group responded honestly under standard instructions of the MMPI-2 manual and the "dissimulator" group faked trying to create the most socially desirable self-image.

Method

The methodological approach followed the typical lines of a quasi-experimental investigation (*post hoc*), since the groups of participants were previously assigned and, at the same time, also more appropriate guidelines and strategies proposed by Santamaría (2012) for simulation research were followed. Data in this research was taken in part from other more complete databases (Jiménez, Sánchez, & Ampudia, 2008; Sánchez, Jiménez, Merino, Ampudia, & Tobón, 2008) in which the MMPI-2 was administered to Spanish population. Special emphasis was made on diagnostic accuracy through a Receiver Operating Characteristic (ROC).

Participants

A total of 587 subjects (280 male and 307 female) considered to be "normal", with no evidence of any psychological disorder, participated in this study and were divided into two groups called "honest" and "dissimulator". The "honest" group, which completed the MMPI-2 honestly under the manual guidelines, consisted of 309 individuals (163 male and 146 female), mean age of male participants was 32.29 (*SD* = 12.37) and of female participants, 32.57 (*SD* = 11.67).

The "dissimulator" group consisted of 278 subjects (117 male and 161 female), mean age of male participants was 28.08 (*SD* = 9.47) and of female participants, 26.64 (*SD* = 8.12). The latter was specifically instructed to complete the MMPI-2 in a "socially desirable" direction: "You have been given a true or false questionnaire. You must project at all times a socially favorable self-image so as to get a good prize." Participants' level of education, occupation, religion, ethnicity, or socio-cultural status was not taken into account, as none of these variables was considered to be especially influential in the results. Every participant's area of residence is in one of the many regional communities of the Spanish territory. Protocols with a determined number of items left unanswered (item ≥ 15) or showing response inconsistency (VRIN-r or TRIN-r > 79T) were excluded (Ben-Porath & Tellegen, 2008; Ben-Porath, 2012). All participants responded voluntarily and in a disinterested manner.

Instruments

Data provided by Ben-Porath and Tellegen (2008) in the MMPI-2-RF are results from the same items (567) obtained through the administration of the MMPI-2 with the relevant reduction down to 338 items (Ben-Porath, 2012).

The material used for this study is the Spanish adaptation of the Minnesota Multiphasic Personality Inventory-2 (MMPI-2, Ávila & Jiménez, 1999; Butcher, Dahlstrom, Graham, Tellegen, & Kaemer, 1989). This study starts from items initially composing the Esd scale of Edwards (37), the Other Deception scale ODecp (33), the *Wsd* scale of Wiggins (33), and the S scale of Butcher and Han (50) in the MMPI-2 (Greene, 2000). Later on, these items were adapted to the MMPI-2-RF, after the restructuration made by Ben-Porath and Tellegen (2008) and Santamaría (2009), adding an "r" to their acronyms, as in Esd-r, Odecp-r, Wsd-r, and S-r respectively.

The first version of the Esd Social Desirability scale of Edwards (1953) was composed of 79 items. As part of its improvement process years later, Edwards (1957) conducted a study with ten judges in order to select those MMPI items that evoked socially desirable responses and they selected the items that distinguished individuals with high scores from individuals with low scores on this scale. The scale was reduced down to 39 items, 12 of them matched the F validity scale (Infrequency) and 9 of them matched the A scale (Anxiety) of Welsh and had to be endorsed "false" in order to be considered "socially desirable" (Ben-Porath, 2012). Then, a new restructuration made by Greene (2000) reduced the number of scale items to 37. Generally, this new readjustment reflects "absence of psychopathological problems, good attention and concentration skills, and acceptable social relationships" (p. 102), thus solving the problem of psychopathological symptom load, one of the most frequent criticism that this type of scales attempting to assess social desirability has received (Crowne & Marlowe, 1960; Edwards & Edwards, 1992; Ferrando & Chico, 2000). With the MMPI-2 restructuration (MMPI-2-RF) the scale was reduced down to a total of 24 items and renamed Esd-r and it shares only one item (4.16%) with the K scale.

Wiggins (1959) developed the Wsd Social Desirability scale with the purpose of discriminating subjects (*n* = 178) instructed to complete the MMPI under socially "desirable" conditions from the other subject group (*n* = 140) instructed to respond honestly under standard instructions as described in the MMPI manual. Baer, Wetter, Nichols, Greene, and Berry (1995) found that the Wsd scale added complementarity to the L (Lie) Validity scale and the K (Defensiveness) scale in differentiating students instructed to project a favorable self-image from those who completed the MMPI-2 honestly. In spite of the existing evidence against Wiggins scale, Graham (2000) suggested that it is a good scale, worth including in the "second generation of validity scales" battery in the MMPI-2. With the MMPI-2 restructuration (MMPI-2-RF), this scale was renamed Wsd-r and

reduced down to 14 items in total, 12 of them are shared with the ODecp-r scale (80%) and 7 (50%) of them with the L scale. It does not share any item with the Esd-r scale of Edwards.

The positive malingering or the Other Deception scale, ODecp, of Nichols and Greene (1991) was originally called Positive Malingering (Mp) and developed by Cofer, Chance, and Judson (1949) to identify defensiveness in people with psychopathological disorders who wish to project a favorable self-image (Greene, 2011). To conduct their study, they asked one group of students to endorse the MMPI items as if they were emotionally disturbed (negative malingering) and another one to respond so as to make the best possible impression (positive malingering). Then, they developed a scale with 34 items which could help identify defensive people. Baer, Wetter, and Berry (1992) found in their meta-analysis that Mp was highly sensitive for discriminating students instructed to be defensive and to project a good self-image. The optimal cutting scores of student samples range from +9 (Bagby, Rogers, Buis, & Kalemba, 1994) to +13 (Baer et al., 1995) and +14 (Bagby, Rogers, & Buis, 1994) according to their own research. Nichols & Greene (1991) further developed the Mp scale, renamed it Other Deception (ODecp), and reduced it down to 33 items by combining Mp scale items with Wiggins (1959) Social Desirability (Wsd) scale and eliminating those with a low total-item correlation. These items where shared, in different proportions, across L and K scales, the Superlative (S) scale of Butcher and Han (1995), and Wiggins Social Desirability (Wsd) Scale, essentially reflecting self-trust and assurance of not having any psychological problems. With the restructuration of the MMPI-2 (MMPI-2-RF) the positive malingering scale was renamed Odecp-r and reduced down to 17 items, 12 of them (70.6%) shared with Wiggins scale (Wsd-r) and 1 of them (5.8%) with L scale.

The Superlative scale of Butcher and Han (2005) was developed to assess individuals who present a self-image of exaggerated virtues and minimized or concealed faults. The research started with a pilot study composed of 274 MMPI-2 items administered to 1138 persons, being reduced to a total of 52 items to establish, later on with an homogeneity analysis, 50 items (Greene, 2000). With the MMPI-2 restructuration (MMPI-2-RF), the Superlative (S-r) scale was reduced down to 26 items and now shares 1 item (3.85%) with Edwards scale, 2 items (7.70%) with the Other Deception scale, 3 items (11.54%) with the Adjustment Validity (K-r) scale, and none with the Uncommon Virtues (L-r) scale nor with Wiggins scale.

Data analysis

Based on the main purpose of this study, the following scales were analyzed generally and comparatively: Edwards (Esd-r) scale, the scale of Nichols & Greene (ODecp-r), Wiggins (Wsd-r) scale, and Superlative (S-r) scale. Specifically, a homogeneity and reliability analysis (Cronbach's a) was carried out for each scale; mean scores differences (Student's t) were obtained for the scales in each group, the honest and the dissimulator; their associative relation with the scale battery that triggers minimization of faults (faking good) on the MMPI-2RF (L-r, F-r, and S-r) was analyzed; and, finally, their diagnostic accuracy was determined by the area under curve (AUC), sensitivity, specificity and predictive power, provided by the Receiver Operating Characteristic (ROC) analysis method. The latter analysis was carried out since it is a methodology that was developed within the Decision Theory in the 50's (Swets & Pickett, 1982) and was initially designed to detect radar signals and subsequently applied to biomedicine (Zweig & Campbell, 1993). In the final graphic of both scales (Figure 1) the specific contribution of each can be appreciated more clearly.

Results

By analyzing scores obtained from the different social desirability scales, taking gender into account, it was observed that,

notwithstanding the existence of significant differences between some scales in various levels, the effect size (Cohen's *d*) proved these statistical differences to be "non-important" (see Table 1). No significant differences in scores by mean age or gender between groups were found either. This is the reason why this study was omitted from statistical analyses for *gender* and *age* variables.

Scores obtained by the "honest" and "dissimulator" groups on the different scales were analyzed, as well as mean scores and their statistically significant difference (Student's *t*) in relation to effect size (Cohen's *d*). In every scale, scores are significant, reinforced by Cohen's *d* analysis, as can be seen in Table 2.

In relation to internal consistency (Cronbach's α) of items comprising each scale, it was observed that the dissimulator group showed a higher score (ODecp-r = .748, Wsd-r = .615, Esd-r = .768, and S-r = .825) than the honest group (ODecp-r = .537, Wsd-r = .361, Esd-r = .730, and S-r = .721), with S-r and Esd-r scales showing a slightly higher score for the dissimulator group and the honest group respectively.

Correlations were found between the validity scales studied, as well as their association with scales with the same positive direction and the purpose of minimizing faults (underreporting, faking good, or good image) already included in the MMPI-2-RF: the L-r and K-r scales, and the Superlative (S-r) scale of Butcher & Han (1995). As can be seen in Table 3, the Wsd-r and the ODecp-r scales show high values (dissimulator = .904, honest = .858; the high percentage of item overlapping should be considered). Let us take a look at the ODecp-r (r = .729) and Esd-r (r = .704) values for the dissimulator group as compared with the Uncommon Virtues (L-r) scale, essentially derived from the number of overlapping items.

By looking at the comparative scale analysis (Table 4), three fundamental characteristics can be observed: 1) Edwards's (Esd-r) scale appears to be the one with the lowest diagnostic accuracy results globally; 2) the other two scales, Other Deception (ODecp-r) and Wiggins (Wsd-r) drew very similar global results, basically because of the high item overlapping (12 out of 17 items of the ODecp-r are common to the Wsd-r, 70.6%) between them; and 3) both the ODecp-r and Wsd scales or the L-r scale do not show diagnostic accuracy. These differences are statistically significant and to be considered.

Among typical characteristics of diagnostic accuracy, it is important that both Positive Predictive Power (PP+) and Negative Predictive Power (PP-) at different incidence levels be noted. Since prevalence of dissimulation rates is unknown, this study analyzed three levels: 15, 20, and 30%. Results show two interesting notions: 1) the higher the prevalence rate, the higher the Positive Predictive Power (PP+) and the lower the Negative one (PP-); and 2) the Odecp-r scale shows the best predictive power, followed by the L-r and Wsd-r scales.

In Table 6, the results obtained from statistical differences on the areas under the curve (AUC) are presented. It is worth noting that there are no statistically significant differences between the ODecp-r, Wsd-r, and L-r scales. Edwards's (Esd-r) scale shows significant differences in diagnostic accuracy (AUC) compared to the other scales.

Discussion and conclusion

Researchers' concern and interest in detecting social desirability bias on diverse assessment techniques is previous even to the MMPI

Table 1
Statistical gender differences (Student's *t*) between scales and their relation to effect size (Cohen's *d*)

Group	Scales	Gender	N	Mean	SD	Sign. *t*	Cohen's *d*
Honest	ODecp-r	Male	163	8.46	2.765	.001	0.200
		Female	146	7.40	2.569		
	Wsd-r	Male	163	6.63	2.135	.015	0.140
		Female	146	6.04	2.071		
	Esd-r	Male	163	16.79	3.901	.004	0.160
		Female	146	15.55	3.595		
	S-r	Male	163	10.88	4.407	.244	0.010
		Female	146	10.97	4.035		
	L-r	Male	163	5.28	1.669	.801	0.000
		Female	146	5.28	1.601		
	K-r	Male	163	6.55	2.747	.009	0.084
		Female	146	6.12	2.303		
Dissimulator	ODecp-r	Male	117	14.18	2.483	.217	0.070
		Female	161	13.76	3.016		
	Wsd-r	Male	117	10.81	2.165	.653	0.020
		Female	161	10.69	2.295		
	Esd-r	Male	117	20.54	3.013	.654	0.030
		Female	161	20.37	3.255		
	S-r	Male	117	17.62	4.785	.996	0.053
		Female	161	18.13	4.638		
	L-r	Male	117	8.89	2.012	.800	0.059
		Female	161	9.14	2.084		
	K-r	Male	117	10.11	2.188	.445	0.053
		Female	161	9.86	2.457		

Note. ODecp-r: restructured Other Deception scale of Nichols & Greene (1991); Wsd-r: restructured Social Desirability scale of Wiggins (1959); Esd-r: restructured Social Desirability scale of Edwards (1957); L-r: restructured Uncommon Virtues; K-r: restructured Adjustment Validity; S-r: restructured Superlative scale of Butcher & Han (1995).

Table 2
Mean score differences between scales and their relation to effect size (Cohen's *d*)

Scale	Group	Mean	Standard deviation	Mean Differences (Student's *t*)	Cohen's *d*
ODecp-r	Honest	7.96	2.722	-5.98*	1.08
	Dissimulator	13.94	2.807		
Wsd-r	Honest	6.35	2.122	-4.39*	1.00
	Dissimulator	10.74	2.238		
Esd-r	Honest	16.20	3.804	-4.24*	0.60
	Dissimulator	20.44	3.151		
S-r	Honest	10.92	4.229	-7.00*	0.78
	Dissimulator	17.92	4.699		
L-r	Honest	5.28	1.634	-3.75*	1.05
	Dissimulator	9.03	2.054		
K-r	Honest	6.34	2.552	-3.62*	0.74
	Dissimulator	9.96	2.347		

Note. *Mean differences are significant at .05. ODecp-r: restructured Other Deception scale of Nichols & Greene (1991); Wsd-r: restructured Social Desirability scale of Wiggins (1959); Esd-r: restructured Social Desirability scale of Edwards (1957); L-r: restructured Uncommon Virtues; K-r: restructured Adjustment Validity; S-r: restructured Superlative scale of Butcher and Han (1995).

Table 3
Correlations between Social Desirability scales and specific *fault-minimizing* scales (*fake good*) included in the MMPI-2-RF

Scale	Group	L-r	K-r	S-r	ODecp-r	Wsd-r
ODecp-r	Dissimulator	.729**	.594**	.644**	1	-
	Honest	.561**	.430**	.294**	1	-
Wsd-r	Dissimulator	.704**	.465**	.545**	.904**	1
	Honest	.551**	.209**	.073	.858**	1
Esd-r	Dissimulator	.527**	.645**	.697**	.661**	.490**
	Honest	.262**	.559**	.513**	.455**	.261**

Note. **Correlation is significant at .01 (bilateral). ODecp-r: restructured Other Deception scale of Nichols & Greene (1991); Wsd-r: restructured Social Desirability scale of Wiggins (1959); Esd-r: restructured Social Desirability scale of Edwards (1957); L-r: restructured Uncommon Virtues; K-r: restructured Adjustment Validity; S-r: restructured Superlative scale of Butcher and Han (1995).

Table 4
Comparative analysis of essential statistics on scale diagnostic accuracy

Scale	AUC	CI	COP	Sensitivity %	Specificity %
ODecp-r	0.922*	0.897-0.942	> 11	82.4	90.6
Wsd-r	0.911*	0.885-0.933	> 8	85.6	85.4
Esd-r	0.820*	0.787-0.851	> 18	81.3	70.6
L-r	0.918*	0.893-0.939	> 7	79.1	92.2
K-r	0.848*	0.816-0.876	> 8	79.1	79.3
S-r	0.859*	0.828-0.886	> 13	83.1	73.1

Note. *Significance (*p* < .0001); AUC = Area under the Curve; CI = Confidence Interval; COP = Cut-Off Point; ODecp-r: restructured Other Deception scale of Nichols & Greene (1991); Wsd-r: restructured Social Desirability scale of Wiggins (1959); Esd-r: restructured Social Desirability scale of Edwards (1957); L-r: restructured Uncommon Virtues; K-r: restructured Adjustment Validity; S-r: restructured Superlative scale of Butcher and Han (1995).

development (Bernreuter, 1933) and reached its peak between 1955 and 1967 (Crowne & Marlowe, 1960; Edwards, 1953, 1957), reviving since the early 90's (Ferrando & Chico, 2000; Jiménez, Sánchez, & Ampudia, 2008; Jiménez, Sánchez, & Tobón, 2009). Some authors dedicated to establish reliability (Kline, 1986; Nunnally, 1987) consider that an empirical test indicating that that test is not affected by the social desirability bias variable should always be included in the validation process for any personality test (Ferrando & Chico, 2000). As the MMPI-2 and its restructured version (MMPI-2-RF) are personality assessment techniques which could be easily manipulated due to their self-report structure, it was just a matter of time before

the personal component of social desirability bias was studied and proposed as another variable in protocol validity. This study has allowed us to analyze, compare, and propose the scale showing more global diagnostic accuracy in the MMPI-2-RF test, recently adapted to Spanish (Santamaría, 2009).

The MMPI-2 technique and its restructured form (MMPI-2-RF) contain valid simulator-detecting instruments, both for symptomathology exaggeration and minimization, i.e., if the MMPI-2-RF already includes scales such as the Uncommon Virtues (L-r) and the Adjustment Validity (K-r) scales, which attempt to detect people trying to minimize their own faults (*underreporting, faking good*), was

Table 5
Positive and Negative Predictive Power of scales based on different prevalence rates

Scale	COP	PP+			PP-		
		Prv.15%	Prv.20%	Prv.30%	Prv.15%	Prv.20%	Prv.30%
ODecp-r	>11	60.8	68.7	79.0	96.7	95.4	92.3
Wsd-r	>8	50.9	59.5	71.6	97.1	96.0	93.3
Esd-r	>18	32.8	40.8	54.2	95.5	93.8	89.8
L-r	>7	58.6	66.6	81.4	96.3	94.9	91.2
K-r	>13	35.3	43.6	57.0	96.1	94.5	91.0
S-r	>8	40.3	48.9	62.1	95.6	93.8	89.9

Note. COP = Cut-Off Point; PP+ = Positive Predictive Power; PP- = Negative Predictive Power; ODecp-r: restructured Other Deception scale of Nichols & Greene (1991); Wsd-r: restructured Social Desirability scale of Wiggins (1959); Esd-r: restructured Social Desirability scale of Edwards (1957); L-r: restructured Uncommon Virtues; K-r: restructured Adjustment Validity; S-r: restructured Superlative scale of Butcher & Han (1995); Prv. = Prevalence.

Table 6
Comparative analysis of differences between scales. Significant differences in the area under the curve (AUC)

Scale	Difference between areas	Standard error	CI (95%)	Signif.
ODecp-r	Esd-r			
	0.101	0.0144	0.0729-0.129	$p < .0001$
	Wsd-r			
	0.0106	0.00595	-0.00107-0.0222	$p = .0751$
	L-r			
	0.00323	0.00978	-0.0159-0.0224	$p = .7413$
	K-r			
	0.0739	0.0137	0.0470-0.102	$p < .0001$
	S-r			
	0.0625	0.0134	0.0362-0.0888	$p < .0001$
Esd-r	Wsd-r			
	0.0906	0.0168	0.0577-0.123	$p < .0001$
	L-r			
	0.0904	0.0176	0.0560-0.125	$p < .0001$
	K-r			
	0.0273	0.0144	-0000998-0.0566	$p = .0586$
	S-r			
	0.0387	0.0148	0.00968-0.0677	$p = .0089$
Wsd-r	L-r			
	0.000157	0.0123	-0.0239-0.0243	$p = .9898$
	K-r			
	0.0633	0.0161	0.0317-0.0949	$p < .0001$
	S-r			
	0.0519	0.0154	0.0217-0.0821	$p = .0007$
L-r	K-r			
	0.0631	0.0159	0.321-0.0942	$p < .0001$
	S-r			
	0.0518	0.0143	0.0236-0.0799	$p = .0003$
K-r	S-r			
	0.0114	0.0106	-0.00940-0.0322	$p = .2831$

Note. CI: Confidence Interval; ODecp-r: restructured Other Deception scale of Nichols & Greene (1991); Wsd-r: restructured Social Desirability scale of Wiggins (1959); Esd-r: restructured Social Desirability scale of Edwards (1957); L-r: restructured Uncommon Virtues; K-r: restructured Adjustment Validity; S-r: restructured Superlative scale of Butcher and Han (1995).

it really necessary a social desirability bias scale for the MMPI-2-RF or does it already exist? This was the question asked by the authors of this study, who nevertheless decided to conduct a research for three essential reasons: a) social desirability is an inherent personality characteristic of all individuals and personality assessment techniques should include its detection; b) could there exist a MMPI-2-RF variable for detecting social desirability inherent characteristics?; and c) the fact that in the Minnesota Multiphasic
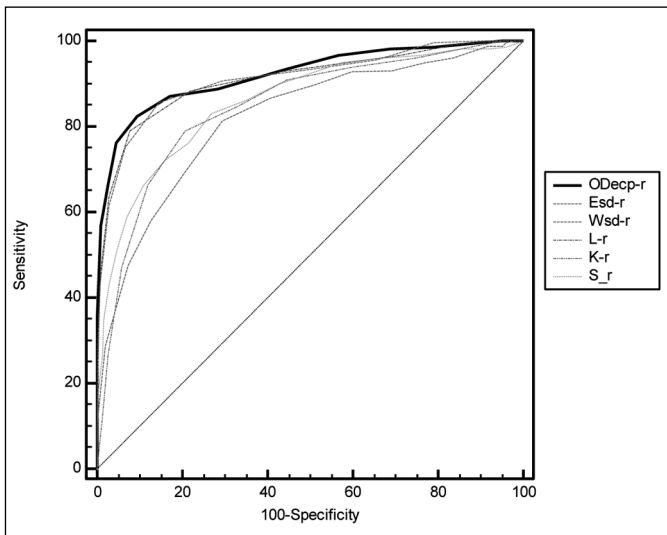
Figure 1. Graphic representation of Sensitivity and Specificity of the three analyzed scales.

Personality Inventory (MMPI) reference is already made to research on social desirability bias conducted by previous authors (Edwards, 1957; Nichols & Greene, 1991; Wiggins, 1959).

With the ROC analysis, similarities and differences on diagnostic accuracy were settled (Area under the Curve, Sensitivity, Specificity, Positive and Negative Predictive Power). Also with the aforesaid analysis, the high diagnostic accuracy of the Wsd-r (Wiggins 1959) and the ODecp-r (Nichols & Greene, 1991) scales, though very similar due mainly to their high item overlapping (70.6% of the ODecp-r items are the same in the Wsd-r scale), was demonstrated (AUC, Table 4). Furthermore, the Uncommon Virtues (L-r) scale, already included in the MMPI-2-RF, was found to offer very similar values on diagnostic accuracy –with no statistically significant differences in the Area Under the Curve (AUC)– to those of the Wiggins Social Desirability (Wsd-r) and Deception (ODecp-r) scales. In addition, data provided by this study has demonstrated that Edwards's (Esd-r) scale show the lowest accuracy in most results.

Diagnostic accuracy is not only determined by AUC value ("Area under the curve": the probability of classifying correctly a couple of dissimulator/honest individuals chosen randomly, based on the results from administering a diagnostic test) but also by Sensitivity, Specificity, and both Positive (PP+) and Negative (PP-) Predictive Power. We already know that Sensitivity and Specificity ROC curve values vary depending on selected *cut-off points*. The ones in this study (Table 4) are the ones which, considering both values, show a better balance in relation to the aforementioned *cut-off point*. Sensitivity is considered as the probability of classifying correctly an individual whose actual condition is defined as "positive" (dissimulator). And Specificity is the probability of classifying correctly an individual whose actual condition is defined as "negative" (honest). Moreover, Nichols and Greene's (ODecp-r) scale shows a probability of 92.2% (AUC) of classifying correctly a couple of dissimulator/honest individuals when chosen randomly. Sensitivity indicates that we can classify correctly a dissimulator individual as such in 82.4% of assessed subjects, but the risk of being mistaken and a false negative is 17.6%. Similarly, with this scale, the probability of diagnosing a person as honest is 90.6%, but the risk we run of considering it as a false positive is 9.4%. The Uncommon Virtues scale (L-r) has obtained similar results, with an AUC = 0.918, Sensitivity of 79.1%, and Specificity of 92.3% at cut-off point > 7. The Wiggins (Wsd-r) scale, with similar results, is the third scale that, along with the other two scales mentioned, shows no statistically significant differences in diagnostic accuracy (Table 6), but there is the item overlapping problem, as we mentioned before.

Both Sensitivity and Specificity provide information about the probability of obtaining a specific result (positive or negative) depending on the actual dissimulator or non-dissimulator condition. However, if the test result is positive, or negative, we should ask ourselves what is the probability that this person is dissimulating. The Predictive Power or Value test can answer that. Positive Predictive Power (PP+) is considered as the probability of being dissimulator when a positive scale result is obtained. Correspondingly, Negative Predictive Power is the probability of being honest when a negative scale result is obtained. We are aware that these predictive values depend basically on dissimulation prevalence and incidence. Taking analyses by other researchers (Sellbom & Bagby, 2010; Wygant et al., 2011) into account and due to the lack of information about specific percentage of incidence in our country, three rates have been proposed: 15%, 20%, and 30% (Table 5). This piece of information about Predictive Power, added to data provided by the Area under the Curve (AUC), will allow us to determine more clearly the diagnostic accuracy for each analyzed scale.

It is evident that working with high sensitivity and specificity diagnostic scales would be ideal, but this is not always possible. A very specific test would be especially appropriate in legal and disease simulation fields, thus avoiding false positives, even with high sensitivity so as to not being at risk of diagnosing incorrectly an actual dissimulator subject as honest. It should be borne in mind that simulation, unlike one-dimensional concepts (González et al., 2012), cannot be the only target of a given scale regardless of how high its diagnostic accuracy is. So, a multidimensional analysis with multiple diagnostic criteria would be more appropriate (Slick, Sherman, & Iverson, 1999).

When essential diagnostic accuracy results in this study (on the MMPI-2-RF) are compared to those of one carried out by Jiménez et al. (2008) –who deliberately and coherently provide their *good image* to others– on the MMPI-2 (*n* = 278), the *ODecp-r* scale (ODecp, with 33 items back then) of Nichols & Greene (1991) throws light on two interesting aspects: on the one hand, significant increased internal consistency (Cronbach's α) for items comprising the MMPI-2-RF (from 0.458 up to 0.748) and on the other hand, a high similarity in global diagnostic accuracy in the area under the curve (AUC), sensitivity, specificity, and predictive power (both positive and negative) values, all of them from 80-90%.

Research by Jiménez et al. (2009) on the MMPI-2 offered a comparative analysis between Wsd (Wiggins, 1959) and Esd (Edwards, 1953) scales, showing diagnostic accuracy, sensitivity, and specificity for each one of them, based on data collected by means of a Receiver Operating Characteristic (ROC) analysis. So by choosing one of them and based on obtained results, authors favored Wiggins's Wsd scale.

At this point, considering our initial hypothesis, "Is there any Social Desirability scale in the MMPI-2-RF?", our conclusion must be affirmative, although it is not identified as a "social desirability" bias detecting scale by name. Which is it? The Uncommon Virtues (*L-r*) scale, included in the validity scale battery of the MMPI-2-RF (Santamaría, 2009), is the one which has proved to have essential criteria for diagnostic accuracy.

Summing up, that is to say that by comparing the results from these six scales in the MMPI-2-RF (ODecp-r, Wsd-r, Esd-r, L-r, S-r, and K-r), three conclusions can be drawn: 1) the MMPI-2-RF includes in its validity scale battery a scale called Uncommon Virtues (L-r), which offers very similar global values on diagnostic accuracy, and no statistically significant differences, to the scale of Nichols & Greene (ODecp-r); 2) the latter scale, ODecp-r of Nichols & Greene (1991) is the one showing the best values on diagnostic accuracy for detecting individuals who present themselves as socially desirable; and 3) we reject the results obtained from Wiggins's (Wsd-r) scale due to its high item overlapping with the ODecp-r scale.

This study does certainly leave pending a series of possibilities that can be subject of future research: working with actual samples,

e.g., personal selection situations (Salgado, 2005), individuals involved in litigations, people with somatic issues, psychological disorders, etc., so as to provide a truer vision of research on these scales. Similarly, there are important limitations along with the risk of not detecting false positives or negatives when we have established a series of criteria related to diagnostic accuracy. Therefore, social desirability bias assessment should be focused, once more, on multidimensional analysis and not only on a specific scale. Ultimately, its applicability will depend mainly on the future use of the present adaptation.

Conflicts of interest

The authors of this article declare no conflicts of interest.

References

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders (DSM-IV-TR)*. Washington, DC: APA.

Andrews, P., & Meyer, G. (2003). Marlowe-Crowne Social Desirability Scale and short Form C: Forensic norms. *Journal of Clinical Psychology, 59*, 483-492. doi: 10.1002/jclp.10136

Arce, R, Fariña, F., Carballal, A., & Novo, M. (2006). Evaluación del daño moral en accidentes de tráfico. Desarrollo y validación de un protocolo para la detección de simulación. *Psicothema, 18*, 278-284.

Archer, R. P., Buffington-Vollum, J. K., Vauter-Strendy, R., & Handel, R. W. (2006). A survey of psychological test use patterns among forensic psychologists. *Journal of Personality Assessment, 87*, 84-94.

Ávila, A., & Jiménez, F. (1999). *Adaptación española del MMPI-2* [*Manual for the reestandaridized Minnesota Multiphasic Personality Inventory: MMPI-2*]. Madrid: TEA Ediciones S. A.

Baer, R. A., Wetter, M. W., & Berry, D. T. R. (1992). Detection of underreporting of psychopathology on the MMPI: A meta-analysis. *Clinical Psychology Review, 12*, 509-525. doi: 10.1016/0272-7358(92)90069-K

Baer, R. A., Wetter, M. W., Nichols, D.S., Greene, R., & Berry, D. T. R. (1995). Sensitivity of MMPI-2 validity scales to underreporting of symptoms. *Psychological Assessment, 7*, 419-423. doi: 10.1037/1040-3590.7.4.419

Bagby, R. M., Rogers, R., & Buis, T. (1994). Detecting malingered and defensive responding on the MMPI-2 in a forensic inpatient sample. *Journal of Personality Assessment, 62*, 191-203. doi: 10.1207/s15327752jpa6202_2

Bagby, R. M., Rogers, R., Buis, T., & Kalemba, V. (1994). Malingered and defensive response styles on the MMPI-2: An examination of validity scales. *Assessment, 1*, 31-38. doi: 10.1177/1073191194001001005

Bagby, R. M., & Marshall, M. B. (2004). Assessing underreporting response bias on the MMPI-2. *Assessment, 11*, 115-126. doi: 10.1177/1073191104265918

Barrick, M. R., & Mount, M. K. (1996). Effects of impression management and self-deception on the predictive validity o personality constructs. *Journal of Applied Psychology, 81*, 261-272.

Ben-Porath, Y. S., & Tellegen, A. (2008). *MMPI-2 Restructured Form: Manual for administration, scoring and interpretation*. Minneapolis: University of Minnesota Press.

Ben-Porath, Y. S. (2012). *Interpreting the MMPI-2-RF*. Minneapolis: University of Minnesota Press.

Bernreuter, R. G. (1933). Validity of the personality inventory. *Personality Journal*, 11, 383-386.

Borum, R., & Grisso, T. (1995). Psychological test use in criminal forensic evaluations. Professional Psychology: *Research and Practice, 26*, 465-473.

Butcher, J., Dahlstrom, W., Graham, J., Tellegen, A., & Kaemmer, B. (1989). *Manual for the reestandaridized Minnesota Multiphasic Personality Inventory: MMPI-2*. Minneapolis: University of Minnesota Press.

Butcher, J. N., & Han, K. (1995). Development of an MMPI-2 scale to assess the presentation of self in a superlative manner: The S scale. In J. N. Butcher & C.D. Spielberger (Eds.), *Advances in personality assessment* (vol. 10, pp. 25-50). Hillsdale, NJ: LEA Press.

Butcher, J. N., Graham, J., Tellegen, A., Dalhstron, W., & Kaemmer, B. (2001). *Minnesota Multiphasic Personality Inventory-2 (MMPI-2): Manual for administration, scoring and interpretation (Revised edition)*. Minneapolis: University of Minnesota Press.

Cofer, C. N., Chance, J., & Judson, A. J. (1949). A study of malingering on the MMPI. *Journal of Psychology, 27*, 491-499.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Clinical Psychology, 24*, 349-354. doi: 10.1037/h0047358

Edwards, A. L., & Edwards, L. K. (1992). Social Desirability and Wiggins's MMPI Content Scales. *Journal of Personality and Social Psychology, 62*, 147-153.

Edwards, A. L. (1953). *Manual for the Edwards Personal Preference Schedule*. New York: Psychological Corporations.

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research*. New York: Dryden.

Edwards, A.L. (1962). Social desirability and expected means on MMPI scales. *Educational and Psychological Measurement, 22*(1), 71-76.

Elvekrog, M., & Vestre, N. (1963). The Edwards Social Desirability scale as a short form of the MMPI. *Journal of Consulting Psychology, 27*, 503-507.

Eysenck, H. J., & Eysenck, S. B. G. (1976). *Psychoticism as a dimension of personality*. New York: Crane, Russak & Company.

Ferrando, P., & Chico, E. (2000). Adaptación y análisis psicométrico de la Escala de Deseabilidad Social de Marlowe y Crowne. *Psicothema, 12*, 383-389.

Fordyce, W.E. (1956). Social Desirability in the MMPI. *Journal of Consulting Psychology, 20*, 171-175. doi: 10.1037/h0048547

González, H., Santamaría, P., & Capilla, P. (2012). *Estrategias de detección de la simulación. Un manual clínico multidisciplinar*. Madrid: TEA Ediciones S. A.

Graham, J. (2000). *MMPI-2, Assessing Personality and Psychopathology* (Third Edition). New York: Oxford University Press.

Graham, J, Watts, D., & Timbrook, R. (1991). Detecting fake-good and fake-bad MMPI-2 profiles. *Journal of Personality Assessment, 57*, 264-277. doi: 10.1207/s15327752jpa5702_6

Greene, R. L. (2000). *The MMPI-2. An interpretative manual* (2nd Ed.). Boston: Allyn and Bacon.

Greene, R. L. (2011). *The MMPI-2 / MMPI-2-RF. An interpretative manual* (3rd Ed.). Boston: Allyn and Bacon.

Hanley, C. (1956). Social desirability and responses to items on three MMPI scales: D, Sc and K. *Journal of Applied Psychology, 40*, 324-328. doi: 10.1037/h0041788

Heilbrun, A. (1964). Social-Learning theory, Social desirability and the MMPI. *Psychological Bulletin, 61*, 377-387. doi: 10.1037/h0048620

Jiménez, F., & Sánchez, G. (2003). Fingimiento de la imagen e Índice de Simulación F-K de Gough en la adaptación española del MMPI-2. *Revista de Psicología General y Aplicada, 56*, 305-317.

Jiménez, F., Sánchez, G., & Tobón, C. (2009). A social desirability scale for the MMPI-2. Which of the two: Wiggins (WSD-R) or Edwards (ESD)? *The European Journal of Psychology Applied to Legal Context, 1*(2), 147-163.

Jiménez, F., Sánchez, G., & Ampudia, A. (2008). Utilidad de la escala *Odecp* de Nichols & Greene (1991) en el MMPI-2. *Revista Iberoamericana de Diagnóstico y Evaluación Psicológica, 26*(2), 75-91.

González, H., Santamaría, P., & Capilla, P. (2012). *Estrategias de detección de la simulación. Un manual clínico multidisciplinar*. Madrid: TEA Ediciones S. A.

Kline, P. (1986). *A handbook of test construction*. London: Methuen.

Lees-Haley, P. (1992). Psychodiagnostic test usage by forensic psychologist. *American Journal of Forensic Psychology, 10*, 25-30.

Nichols, D. S., & Greene, F. L. (1991). *New measures for dissimulation on the MMPI/MMPI-2*. Paper presented at the 26th Annual Symposium on Recent Developments in the Use of the MMPI. St. Petersburg Beach, Fl.

Nicholson, R. A., Mouton, G. J., Bagby, R. M., & Buis, T. (1997). Utility of MMPI-2 Indicators of Response Distortion: Receiver Operating Characteristic Analysis. *Psychological Assessment, 9*, 471-479.

Nunnally, J.C. (1987). *Teoría psicométrica*. Mexico: Trillas.

Pelegrina, M., Ruiz-Soler, M.; López, E., & Wallace, A. (2000). Análisis de variables mediante curvas ROC y modelos categóricos. *Psicothema, 12*, 427-430.

Preti, A., Rocchi, M., Sisti, D., Mura, T., Manca, S., Siddi, S., … Masala, C. (2007). The psychometric discriminative properties of the Peters et al.'s Delusions Inventory: a Receiver Operating Characteristic Curve Analysis. *Comprehensive Psychiatry, 48*, 62-69. doi: 10.1016/j.comppsych.2006.05.003

Rogers, R. (2008). *Clinical assessment of malingering and deception* (3rd Ed.). New York: The Guilford Press.

Rogers, R., & Bender, D. (2003). Evaluation of malingering and deception. In A. Goldstein & I. B. Weiner (Eds.), *Handbook of Psychology. Vol. 11: Forensic Psychology* (109-132). New Jersey: John Wiley and Sons Inc.

Salgado, J. (2005). Personalidad y deseabilidad social en contextos organizacionales: implicaciones para la práctica de la Psicología del Trabajo y las Organizaciones. *Papeles del Psicólogo, 92*, 65-75.

Sánchez, G., Jiménez, F., Merino, V., Ampudia, A., & Tobón, C. (2008). Contribución de la escala de Deseabilidad Social de Edwards (So) a la validez del MMPI-2. *Revista de Psicología General y Aplicada, 61*, 203-216.

Santamaría, P. (2009). MMPI-2-RF. *Inventario Multifásico de Personalidad de Minnesota-2 Reestructurado*. Madrid: TEA Ediciones S. A.

Santamaría, P. (2012). Metodología de la investigación en simulación. In H. González, P. Santamaría & P. Capilla (Coords.), *Estrategia de detección de la simulación. Un manual clínico multidisciplinar* (pp.105-150). Madrid: TEA Ediciones S. A.

Sellbom, M., & Bagby, R.M. (2010). Detection of overreported psychopathology with the MMPI-2-RF form validity scales. *Psychological Assessment, 22*, 757-767. doi: 10.1037/a0020825

Slick, D. J., Sherman, E.M., & Iverson, G. L. (1999). Diagnostic criteria for malingering neurocognitive dysfunction: Proposed standards for clinical practice and research. *The Clinical Neuropsychologist, 13*, 545-561. doi: 10.1076/1385-4046(199911)13:04;1-Y;FT545

Swets, J. A., & Pickett, R. M. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory*. New York: Academic Press.

Wiggins, J. S. (1959). Interrelations among the MMPI measures of Dissimulation under standard and social desirability instructions. *Journal of Consulting Psychology, 23*, 419-427.

Wygant, D. B., Anderson, J. L., Sellbom, M., Rapier, J. L., Algeier, L. M., & Granacher, R. P. (2011). Association of the MMPI-2 Restructured Form (MMPI-2-RF) Validity Scales with Structured Malingering Criteria. *Psychological Injury and Law, 4*, 13-23. doi: 10.1007/s12207-011-9098-z

Zweig, M. H. & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry, 39*, 561-577.