ORIGINAL RESEARCH—INTERSEX AND GENDER IDENTITY DISORDERS

Assessing the Utility of Diagnostic Criteria: A Multisite Study on Gender Identity Disorder

Muirne C.S. Paap, MS,*† Baudewijntje P.C. Kreukels, PhD,‡ Peggy T. Cohen-Kettenis, PhD,‡ Hertha Richter-Appelt, PhD,§ Griet de Cuypere, MD, PhD,¶ and Ira R. Haraldsen, MD, PhD*

*Oslo University Hospital—Clinic for Gender Identity Disorder, Department of Neuropsychiatry and Psychosomatic Medicine, Norway; †University of Oslo—Institute of Clinical Medicine, Norway; ‡VU University Medical Center—Department of Medical Psychology, Amsterdam, the Netherlands; §University Hospital Hamburg-Eppendorf—Institute for Sex Research and Forensic Psychiatry, Hamburg, Germany; ¶University Hospital—Department of Sexology and Gender Problems, Gent, Belgium

DOI: 10.1111/j.1743-6109.2010.02066.x

ABSTRACT

Introduction. Studies involving patients with gender identity disorder (GID) are inconsistent with regard to outcomes and often difficult to compare because of the vague descriptions of the diagnostic process. A multisite study is needed to scrutinize the utility and generality of different aspects of the diagnostic criteria for GID.

Aim. To investigate the way in which the diagnosis-specific Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision criteria for GID were used to reach a psychiatric diagnosis in four European countries: the Netherlands (Amsterdam), Norway (Oslo), Germany (Hamburg), and Belgium (Ghent). The main goal was to compare item (symptom) characteristics across countries.

Methods. The current study included all new applicants to the four GID clinics who were seen between January 2007 and March 2009, were at least 16 years of age at their first visit, and had completed the diagnostic assessment (N = 214, mean age = 32 ± 12.2 years). Mokken scale analysis, a form of Nonparametric Item Response Theory (NIRT) was performed.

Main Outcome Measures. Operationalization and quantification of the core criteria A and B resulted in a 23-item score sheet that was filled out by the participating clinicians after they had made a diagnosis.

Results. We found that, when ordering the 23 items according to their means for each country separately, the rank ordering was similar among the four countries for 21 of the items. Furthermore, only one scale emerged, which combined criteria A and B when all data were analyzed together.

Conclusions. Our results indicate that patients' symptoms were interpreted in a similar fashion in all four countries. However, we did not find support for the treatment of A and B as two separate criteria. We recommend the use of NIRT in future studies, especially in studies with small sample sizes and/or with data that show a poor fit to parametric IRT models. **Paap MCS, Kreukels BPC, Cohen-Kettenis PT, Richter-Appelt H, de Cuypere G, and Haraldsen IR. Assessing the utility of diagnostic criteria: A multisite study on gender identity disorder. J Sex Med 2011;8:180–190.**

Key Words. Diagnosis; DSM; Item Response Theory; Gender Identity Disorder; International Study; Transsexualism

Introduction

In the past few decades, sex reassignment surgery as a treatment for transsexualism has been gaining ground; in many countries transsexuals are now being diagnosed and treated by specialists. Scientific interest in the phenomenon of transsexualism [1] or gender identity disorder (GID) [2] has been increasing, which is reflected in a growing body of research on this particular disorder, especially by specialists working in multidisciplinary gender teams [3–12]. So far, research has shown international differences in sex ratio, comorbidity, and sociodemographic variables [6,13]. Differences between subgroups have received much attention in the literature, for example, comparing male-to-female (MtF) with female-to-male (FtM) transsexuals [9,14–16], "early onset" to "late onset" [5,6] of the disorder, or homosexual to heterosexual orientation [4,17]. The effects of cross-sex hormone therapy on cognition have also been extensively studied in many countries [10,11,18,19]. The published results have been far from homogeneous.

One major factor stands in the way of performing a "study of studies" (meta-analysis): the lack of comparability of the data between the publishing clinics and countries [20]. Presently, it is practically impossible to diagnose transsexualism on the basis of objective criteria because of a lack of psychometrically sound psychological instruments to measure the condition [21]. Thus, the next-best thing is a diagnosis made by one or more experienced clinicians. Typically, publications state that GID was diagnosed according to the latest version of the Diagnostic and Statistical Manual of Mental Disorders (DSM) [2] or International Classification of Diseases (ICD) [1] without giving specifics; which makes it impossible to establish whether consensus about a diagnosis would be reached by two clinicians of different clinics. Unfortunately, the criteria as stated in the DSM and ICD leave ample room for interpretation, rendering the reliability of the diagnosis questionable. As far as we know, formal studies investigating the reliability of the diagnosis have not been conducted.

With the increasing popularity of Item Response Theory (IRT), many researchers have started using IRT to explore the utility and generality of diagnostic criteria by carefully scrutinizing symptom criteria and categorization, offering suggestions regarding how to improve upon existing diagnoses on the basis of their results [22–32]. The concept of the "latent trait" plays an important role in IRT. The term latent ("hidden") trait refers to an unobservable variable, which can only be indirectly measured by a set of items. It is the functioning of these items that is central to IRT (as opposed to that of the sum score in Classical Test Theory). An important advantage of IRT is that it allows one to examine the response probability at a particular level of the latent trait; this enables the investigation of item bias or differential item functioning (DIF). DIF is present when two groups of people (for example men and women) have different response probabilities for a particular item, even if they would have *the same score on the latent trait*. Indeed, several researchers have used IRT to demonstrate a link between group membership (gender, psychiatric history) and response probabilities for diagnostic criteria [33,34].

In order to improve comparability of data across clinics as well as diagnostic transparency, we initiated the "European Network for the Investigation of Gender Incongruence" (ENIGI) [35]. The clinics participating in this collaboration are: Amsterdam (the Netherlands), Oslo (Norway), Hamburg (Germany), and Ghent (Belgium). In this collaboration, we chose to use the term "gender incongruence" (GI), which refers to the incongruence between one's gender identity on the one hand, and one's assigned gender and/or one's congenital primary and secondary sex characteristics on the other hand [35,36]. We use GI when referring to those who have not yet been diagnosed with GID or transsexualism.

The main goal of the current study is to investigate the way in which the diagnosis-specific DSM-IV-TR criteria for GID were used to reach a psychiatric diagnosis in the four European countries by using a form of IRT to compare item (symptom) characteristics across countries.

Methods

Subjects

The current study is a part of the ENIGI initiative [35] that includes applicants that were seen at GID clinics in Ghent, Hamburg, Amsterdam, and Oslo from the start of January 2007. The current study included all new applicants that were seen between January 2007 and March 2009, were at least 16 years of age at their first visit, and for whom the diagnostic assessment score sheet had been filled out. Of the 214 included applicants (mean age = 32.3, standard deviation [SD] = 12.2), 89 (41.6%) were natal females (FtMs; mean age = 28.4, SD = 10.4) and 125 (58.4%) were natal males (MtFs; mean age = 35.11,

SD = 12.7). One hundred seventy-six (82.2%) applicants (mean age = 32.8, SD = 12.2) were diagnosed with GID (Ghent: 97.6%, Hamburg: 83.3%, Amsterdam: 88.7%, Oslo: 44.1%), 80 were FtM and 96 were MtF. The low percentage of GID diagnoses in Oslo can partly be explained by a difference in procedure: in Oslo all applicants went through the entire diagnostic phase; as a consequence the diagnostic score sheet was filled out for almost all of them. In the other clinics some applicants were referred elsewhere or dropped out in an early stage of the diagnostic phase. Written informed consent was obtained from the subjects after the study was fully described to them.

Main Outcome Measures

This study focused on the diagnosis-specific criteria A ("strong and persistent cross-gender identification") and B ("persistent discomfort with his or her sex or sense of inappropriateness in the gender role of that sex"). In order to gain more insight into individual clinicians' interpretations of the disorder and the criteria, a scoring sheet was developed which exists of 23 items (Appendix). These items consisted of a combination of a symptom and an "aspect." The aspects were: severity, onset, duration, frequency, persistence. The aspects that were applicable for the given symptom were used. For example, it is noted in the DSM 4th Edition, Text Revision (DSM-IV-TR) [2] that one of the symptoms of the A-criterion is "a stated desire to be the other sex;" we measured this using four items: "how strong," "how persistent," "since when," and "how long." Each item has several scoring possibilities, such as "very strong," "moderately," "mildly" for the first item ("how strong"). For many items, one of the two answering categories "moderately" or "mildly" had very low counts. For this reason, we decided to dichotomize the data: very strong(1)/lesser degree(0), since childhood(1)/later onset(0), longer than 5 years(1)/5 years or shorter(0), very frequently(1)/less frequently(0), very persistent(1)/less persistent(0).

On the whole, the items directly reflect the criteria in the DSM, but the scoring sheet also taps into some aspects of symptoms not currently in the DSM, such as onset and duration. This was done in order to obtain as detailed a picture as possible of the "severity" or "level" of every subcriterium or indicator. The choices were purely made to enable us to describe as accurately as possible what we have been doing until now.

The score sheet was filled out by the participating clinicians after they had made a diagnosis—after approximately seven interviews with the applicant. These seven interviews (each an hour in duration) are both used to diagnose and to determine the "readiness" for treatment. Possible risk-factors and comorbidity are assessed. Some of the patients take a bit longer because they need help with their "coming out." Finally, all potential consequences of treatment are discussed with the patient, as well as medical possibilities and limitations of the treatment, and social consequences of their transformation. This is done to ensure the patient is fully informed before they take the "final step." This diagnostic procedure is highly similar in the four countries [35].

Data Analysis

Mokken scale analysis (MSA) [37] which falls under the category of Nonparametric Item Response Theory (NIRT) [38] was applied to investigate scoring patterns with respect to criteria A and B using the software package MSP5.0 (iecProGAMMA, Groningen, the Netherlands) [39]. MSA both uncovers the dimensionality (factorial structure) of the data, and at the same time provides the researcher with scales that fulfill the criteria of the so-called "Monotone Homogeneity Model" (MHM). This model implies an ordering of *respondents* on an underlying unidimensional scale (measuring GI in our case) using the unweighted sum of item scores [38,40–42].

In addition to the MHM, Mokken [37,43] also proposed the model of double monotonicity (DMM), which allows for the ordering of respondents as well as items on the underlying scale. When the DMM holds, it also implies the same ordering of items in all subgroups, which in our case are "sex" and "clinic," and allows for the investigation of differential item functioning (DIF) or item bias in subgroups [38]. We studied DIF using the information provided by MSP5 on "equal item ordering in subgroups" and subsequently studying the item response functions (IRFs), which depict the relationship between the latent trait and the probability of the item being endorsed. The IRFs were produced by the software package TestGraf 98 (Department of Psychology, McGill University, Montreal, Canada) [44].

Scalability coefficients are the statistics used to check the assumptions underlying the MHM. These coefficients can be calculated between item-pairs (H_{ij}), on item-level (H_i) and on scale-level (H). H_{ij} equals the items' covariance divided by their maximum covariance given their univariate score-frequency distributions [45]. H_i is based on

H_{ij}, and expresses the degree to which an item is related to other items in the scale, comparable—but not identical—with the item-rest correlation in Classical Test Theory (the similarities and differences between H_i and more traditional measures will be discussed more thoroughly in a paper that is currently in progress). Moreover, a high H_i value means that the item distinguishes well between people with relatively low GI values and people with relatively high GI values (comparable with the difficulty parameter in parametric IRT). H is based on H_i and expresses the degree to which the total score accurately orders persons on the GI scale. A scale is considered acceptable if $0.3 \leq H < 0.4$, good if $0.4 \leq H < 0.5$, and strong if $H \geq 0.5$ [37,38].

In the current study, the Algorithm for Item Selection that is available in MSP5 was used to cluster items into a scale, or several scales. Model fit was assessed by checking violations against monotonicity for the MHM, and violations of invariant item ordering (IIO) for the DMM [38]. A monotone scale is one where the participants tend to score higher on items when they have a high GI score, and IIO implies that the more difficult the item, expressed by its smaller mean (probability), the lower the likelihood of endorsing the item given *any* position on the GI scale. There are several methods for checking IIO. We used the rule of thumb proposed by Sijtsma and Meijer [46]: if H^T is larger than or equal to 0.3, and fewer than 10% of the persons have negative H^T_a values, then it is assumed that IIO is not present. H^T compares the score patterns of *0*s and *1*s on all items produced by all individuals. The more similar the item score patterns, the higher H^T. H^T_a is defined on person level, and increases when the score pattern of person *a* is more similar to the average score pattern [38].

Our analysis consisted of three parts: in the first part, the focus was on creating an "international" scale that was valid for all data combined, and in the second part the focus was on item analysis and comparing the item statistics among clinics and between the sexes. Finally, the average scale scores were compared for the clinics and for the sexes by performing Mann–Whitney *U* tests in SPSS 16 (SPSS Inc., Chicago, IL, USA) [47].

Results

Raw Data: Means per Item for Each Country

Table 1 shows the means per item for each country separately, as well as the average item rank per country. As values only range between 0 and 1, these means reflect the proportion of applicants that scored "very strong," "very persistent," "onset in childhood," and "a duration of at least five years." Item ranking was done from low to high for each item, assigning rank "1" to the lowest mean/country, "2" to the second lowest and so on. It can be seen that the average item rank is highest for Ghent, and lowest for Oslo. This indicates that, on average, the means are highest in Ghent and lowest in Oslo. Furthermore, the range of means was quite large, from 0.11 for item *A2_on* ("onset of frequent passing as the other sex") in Amsterdam to 0.95 for item *A3_st* ("strong desire to live or be treated as the other sex") in Ghent.

The range for MtFs was between 0.11 and 0.80; for FtMs it was between 0.28 and 0.95. FtMs scored higher for all items (apart from item *B1_co*; "complete or incomplete preoccupation with getting rid of sex characteristics"); the difference in means between MtFs and FtMs ranged between 0.07 (for item *B1_on*; "onset of preoccupation with getting rid of sex characteristics") and 0.33 (for item *A4_st*; "strong conviction that he or she has the typical feelings of the other sex"). This information is not included in Table 1.

Imputing Missing Data

Missing data occurred for 96 of 4708 cells (2%). Values of "0" (to a lesser degree) were imputed for these cells. Our reasoning was as follows: if the applicant experienced a particular symptom to a high degree, the clinician would have definitely crossed this off as such. Thus, not crossing off anything at least indicated doubt on the part of the clinician, which makes it plausible that the applicant was not experiencing the symptom more strongly than "to a lesser degree." This reasoning was agreed upon by all clinicians involved in this research project.

MSA

Two analyses were carried out: one with "clinic" as the grouping factor and one with "sex" as the grouping factor. Interestingly, we found that patients with an incomplete wish had lower mean scores on all other items than patients with a complete wish (Item B1_co: "complete or incomplete wish with regard to getting rid of the sex characteristics"; indicating, for example, whether the patient is only interested in breast removal/augmentation, or in genital surgery as well). This "perfect prediction" of low vs. high scores led us to remove the item from the MSA, because it had no added value

Table 1 Proportion of applicants scoring "1" on the symptoms of criterion A and B, respectively

Symptom aspect	Item code	Ghent (N = 41)	Hamburg (N = 42)	Amsterdam (N = 97)	Oslo (N = 34)
Criterion A					
Stated desire to be the other sex (A1)					
Strong	*(A1_st)*	0.90	0.79	0.91	0.79
Persistent	*(A1_pe)*	0.83	0.74	0.91	0.68
Onset	*(A1_on)*	0.54	0.50	0.37	0.29
Duration	*(A1_du)*	0.85	0.67	0.67	0.62
Frequent passing as the other sex (A2)					
Often	*(A2_of)*	0.90	0.71	0.84	0.71
Onset	*(A2_on)*	0.37	0.14	0.11	0.15
Duration	*(A2_du)*	0.61	0.40	0.38	0.29
Desire to live or be treated as other sex (A3)					
Strong	*(A3_st)*	0.95	0.74	0.91	0.79
Persistent	*(A3_pe)*	0.83	0.76	0.90	0.74
Onset	*(A3_on)*	0.49	0.26	0.25	0.26
Duration	*(A3_du)*	0.46	0.40	0.54	0.44
Conviction that he or she has the typical feelings of the other sex (A4)					
Strong	*(A4_st)*	0.59	0.76	0.60	0.62
Persistent	*(A4_pe)*	0.68	0.81	0.73	0.59
Onset	*(A4_on)*	0.49	0.40	0.46	0.26
Duration	*(A4_du)*	0.78	0.57	0.78	0.50
Criterion B					
Preoccupation with getting rid of sex characteristics (B1)					
Strong	*(B1_st)*	0.80	0.50	0.73	0.65
Persistent	*(B1_pe)*	0.88	0.57	0.81	0.65
Onset	*(B1_on)*	0.49	0.10	0.13	0.15
Duration	*(B1_du)*	0.78	0.45	0.65	0.56
Complete/incomplete	*(B1_co)*	0.95	0.59	0.86	0.74
Belief to be born the wrong sex (B2)					
Strong	*(B2_st)*	0.76	0.74	0.67	0.65
Onset	*(B2_on)*	0.56	0.38	0.37	0.26
Duration	*(B2_du)*	0.83	0.60	0.72	0.53
Average item rank		3.6	2.0	2.7	1.6

MtF = male-to-female; FtM = female-to-male.

for our scale analyses (and was very different in item content compared with the other items).

Scale Analysis

When all data were analyzed together, only one scale emerged, "the general GI scale." No items were rejected because of negative *H*-values with one of the other scale items and none were excluded because of lower bound and/or significance criteria. H_i-values ranged between 0.45 (item *B1_st;* "strong preoccupation with getting rid of sex characteristics") and 0.74 (item *A2_on;* "onset of frequent passing as the other sex") with $H = 0.55$, indicating that this is a strong, unidimensional scale. Neither the checks for monotonicity nor the checks for IIO ($H^T = 0.48$ and 5% of the applicants showed negative H^T_a values) revealed any deviations when the entire data set was analyzed. Therefore, we assume that the DMM holds.

Item Analysis

The characteristics that were compared between the groups were: number of scales that emerged for each group separately, H_i values, items that were excluded from the scale, violations of monotonicity, violations of invariant item ordering, and equal item ordering in subgroups. A summary of the most important findings of both analyses (comparing the clinics, and the sexes, respectively) can be found in Table 2.

For three of the four clinics, a one-scale solution was found. For Amsterdam, however, two scales emerged from the analysis: one that included the "onset" and "duration" items ("Amst 1") and one that included the "severity" and "persistence" items ("Amst 2"). When all data was divided into two groups based on birth sex, a one-scale solution was found for both the FtM group and the MtF group.

When data were analyzed separately for the subgroups (clinics, sexes), it was found that not all items performed equally well. Two items were excluded from the scale when the Ghent data and the Oslo data were analyzed. In addition, all clinics had at least one item (in Amsterdam there were two) that violated the assumption of monotonicity.

Table 2 Summary of the item analyses for the four clinics and the two sexes

	H (range H_is)	Excluded items	Violations of monotonicity	Violations of invariant item ordering	Violations of equal item ordering in subgroups
Ghent	0.72 (0.59–0.99)	*A3_pe, A3_du*	*A4_on*	—	*A4_st, A4_pe*
Hamburg	0.55 (0.42–0.82)	—	*B1_pe*	—	*A4_st, A4_pe*
Amst 1	0.59 (0.46–0.84)	—	*A2_du*	—	—
Amst 2	0.53 (0.42–0.70)	—	*A4_st*	*A2_of, A4_pe*	*A4_st, A4_pe*
Oslo	0.72 (0.55–0.90)	*A1_on, B1_on*	*A2_du*	—	*A4_st, A4_pe*
FtM	0.53 (0.35–0.79)	*B2_on*	—	—	*B2_st*
MtF	0.51 (0.39–0.74)	*B1_pe*	*A1_pe*	—	*B2_st*

The sex-comparison analyses showed a similar picture. However, no violations were found when all the data were analyzed together.

Overall, the "stronger" Mokken Model, the Double Homogeneity Model (DMM), showed a good fit. When only one scale (the "general GI scale") was allowed for, no violations were found for any of the clinics. The H^T values equaled 0.54, 0.48, 0.50, and 0.45 for Ghent, Hamburg, Amsterdam and Oslo, respectively. The corresponding percentages of persons showing negative H^T_a values were 2.9%, 2.6%, 3.3%, and 5.3%. When the same tests were repeated for the two sexes, no violations were found for either sex (for FtMs, $H^T = 0.55$ and 4.9% had negative H^T_a values for MtFs, $H^T = 0.47$% and 1.7% had negative H^T_a values).

Only two items violated the assumption of equal item ordering in subgroups that is implied by the DMM. Namely, "strong conviction that he or she has the typical feelings of the other sex" (item *A4_st*) and "persistent conviction that he or she has the typical feelings of the other sex" (item *A4_pe*). The IRFs of items *A4_st* and *A4_pe* are depicted in Figures 1 and 2, respectively, for each clinic. If there would have been equal item ordering in subgroups, the lines in these figures would lie on top of each other. Instead, it can be seen from the figures that this is far from the case. Looking at Figure 1, for example, it can be seen that—given the same score on the latent trait (GI in our case)—the probability of scoring "1" on this item is low in Ghent and high in Hamburg, with Amsterdam and Oslo in between. One item violated the DMM model when comparing the sexes: "strong belief to be born the wrong sex" (item *B2_st*). Given the same average score, FtMs had a higher probability of having endorsed this item than MtFs (Figure 3). These findings suggest that items *A4_st*, *A4_pe*, and *B2_st* might be biased (cultural/gender bias) and should be handled with care. We chose not to include these items when calculating an average probability score.

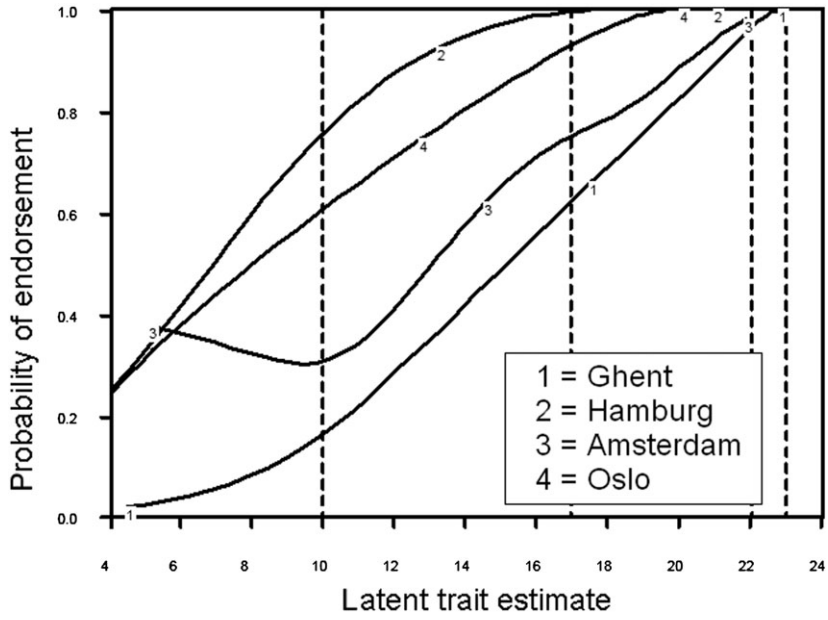

Figure 1 The item response functions (IRFs) of the four clinics for item *A4_st* ("strong conviction that he or she has the typical feelings of the other sex"). Item Response Theory allows for different IRFs to be created for different groups and be placed on a common scale. The IRFs show that patients in Ghent and Amsterdam need to score higher than patients in Hamburg and Oslo on the latent trait estimate for this item to be endorsed.

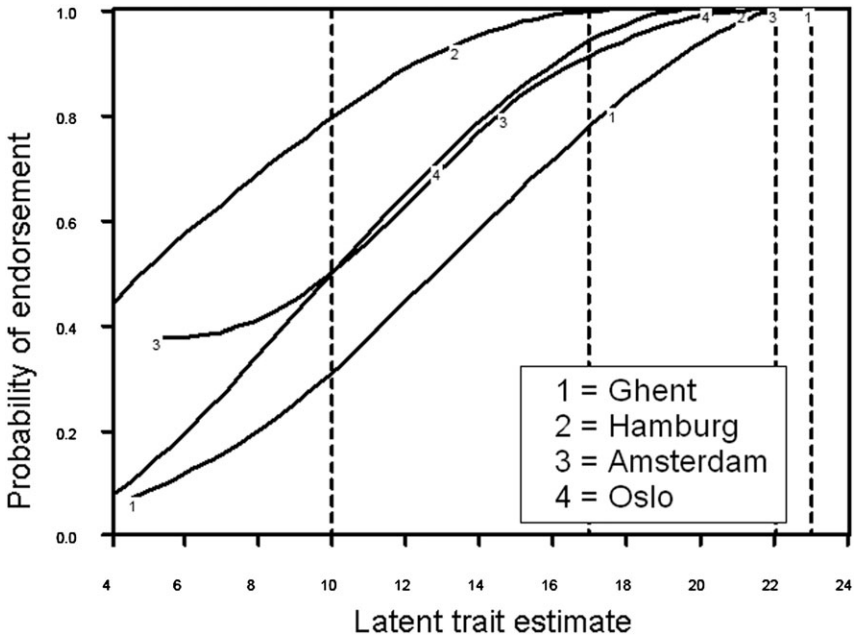

Figure 2 The item response functions (IRFs) of the four clinics for item *A4_pe* ("persistent conviction that he or she has the typical feelings of the other sex"). The IRFs show that patients in Ghent need to score higher than patients in Oslo, Amsterdam, and Hamburg on the latent trait estimate for this item to be endorsed.

Mean Scores: Comparing Distributions

The "average probability" was based on the "general GI scale," omitting items *A4_st*, *A4_pe*, and *B2_st*. The "average probability" was not normally distributed for all subgroups. Comparisons were therefore made by using nonparametric tests. An overview of descriptive statistics for each subgroup can be found in Table 3.

When considering the data of all applicants, regardless of diagnosis, it was found that the medians of Hamburg, Amsterdam, and Oslo were highly comparable. Ghent's median was significantly higher than those of the other clinics (Hamburg: Mann-Whitney U [83] = 538, $P = 0.003$; Amsterdam: Mann-Whitney U [138] = 1508, $P = 0.025$; Oslo: Mann-Whitney U [75] = 411, $P = 0.002$). Comparison of the sexes revealed that FtMs showed a significantly higher median than MtFs (Mann–Whitney U [214] = 3466, $P < 0.001$).

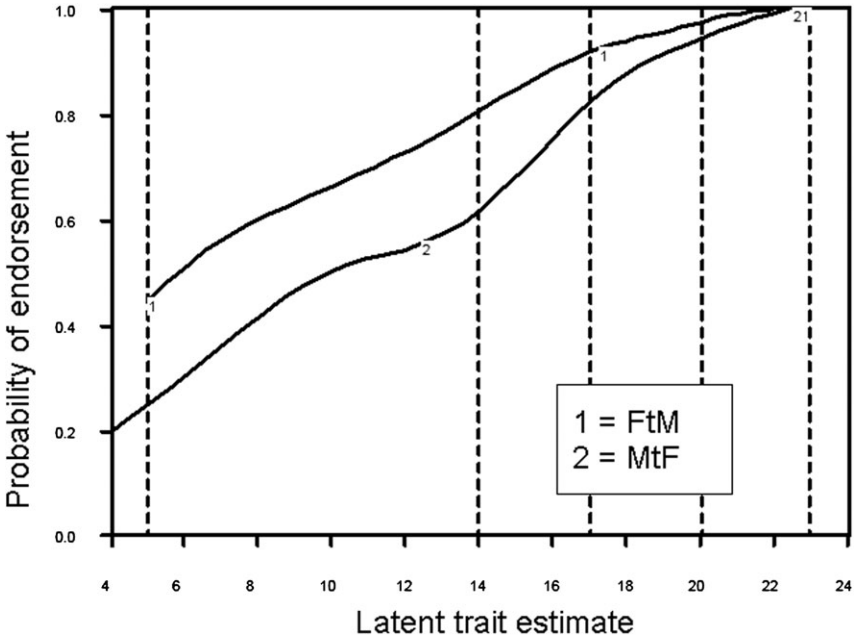

Figure 3 The item response functions (IRFs) showing gender differences for item *B2_st* ("strong belief to be born the wrong sex"). The IRFs show that male-to-females (MtFs) have a lower probability of endorsement than female-to-males (FtMs) at the same level of the latent trait estimate.

Table 3 Median and interquartile range (IQR) for the "average probability" of scoring "1" by subgroups

	Total (N = 214)	FtM (N = 89)	MtF (N = 125)	Ghent (N = 41)	Hamburg (N = 42)	Amsterdam (N = 97)	Oslo (N = 34)
All applicants							
Median	0.63	0.70	0.55	0.70	0.63	0.60	0.58
IQR*	0.40	0.32	0.35	0.52	0.48	0.30	0.60
GID applicants							
Median	0.68	0.75	0.60	0.78	0.65	0.65	0.80
IQR	0.37	0.30	0.34	0.50	0.45	0.25	0.15

*75th percentile–25th percentile.
GID = gender identity disorder; MtF = male-to-female; FtM = female-to-male.

However, when only the data of applicants diagnosed with GID were considered, all medians (save for the Amsterdam and FtM medians) had a higher value than those for the whole group of applicants. The largest increase was seen for Oslo, which now differed from Hamburg (Mann–Whitney U [50] = 154, $P = 0.021$) and Amsterdam (Mann–Whitney U [101] = 404, $P = 0.021$), and this increase in the value of the median was accompanied by a decrease in spread. Ghent's median was significantly higher only than Hamburg's (Mann–Whitney U [75] = 466, $P = 0.013$) and Amsterdam's (Mann-Whitney U [126] = 1345, $P = 0.048$). The difference in medians between the sexes diminished, but remained statistically significant (Mann–Whitney U [176] = 2520, $P < 0.001$).

Discussion

In the current study, MSA was used to evaluate whether the DSM-IV-TR diagnostic criteria for GID were used in a similar fashion in four European clinics, and whether they were used similarly when diagnosing natal males (MtF) and females (FtM). The diagnostic criteria were operationalized and quantified on item-level, and an item-analysis and a scale-analysis were conducted.

Our results showed that the GID criteria themselves were largely interpreted in the same way in the four clinics participating in this study. Most criteria were free of cultural and gender bias. When the data of all clinics were analyzed jointly, only one scale emerged, which comprised the diagnosis-specific criteria A and B (the "general GI scale"). This one-scale solution was also found for three of the clinics (Ghent, Hamburg, Oslo) when the data was analyzed separately for each clinic, and for both sexes when the data was analyzed separately for each sex. In Amsterdam, a two-scale solution was found: one scale consisted of all duration and onset items, and the other scale consisted of all strength and persistence items. A new study is needed to explain the difference in scale solutions between Amsterdam and the other clinics. A possible explanation could be that Dutch patients present themselves differently than other patients. It could, however, also mean that Dutch clinicians have a different way of diagnosing. The difference in scale solutions might also lead to differences in diagnostic decisions; in Amsterdam, an applicant could still receive the diagnosis when symptoms are very severe and persistent but of relatively recent onset, whereas this seems less likely to happen in the other clinics. In spite of relatively low means for the onset items, these items showed a strong relationship with the other items in Ghent, Hamburg and Oslo. At this point we cannot say whether the strong relationship of onset items with the other items in the scale is caused by a "real" strong relationship between symptoms experienced by the patients or caused by the frame of reference of the clinicians in these clinics. The differences between Amsterdam and the other clinics illustrate the importance of a multisite study when scrutinizing the usefulness of diagnostic criteria, and of fuelling the ongoing discussion of the generality and standardization of diagnoses in cross-cultural settings. Our findings indicate that the subdivision into two criteria (A and B) that was introduced in the DSM-IV is likely to be superfluous.

Moreover, our results indicate that the link between the GI score and subsequent diagnosis might differ among the participating clinics. We found that diagnostic thresholds differ among the clinics: Ghent and Oslo have higher thresholds for GID than Amsterdam and Hamburg. However, the median for all applicants regardless of diagnosis was much higher in Ghent than in Oslo; thus, it is unclear whether the high threshold in Ghent is attributable to a referral bias of applicants in Flanders, or whether it reflects a systematic difference in judgment between clinicians in Ghent and Oslo. The biggest shift in medians (and spread) was seen in Oslo, when the total group was com-

pared with the GID group: the GID group scored rather high on the "general GI scale," compared with the total group. This, in combination with the observation that only 44.1% of the total patient group received the diagnosis, vs. 83.3%–97.6% in the other clinics, could reflect a more "conservative" view of GID in Oslo, and the low spread in scores for applicants diagnosed with GID in Oslo could reflect a narrower interpretation of the GID criteria than in the other clinics. However, in Oslo all applicants went through the first part of the diagnostic phase (6 months) and as a consequence the diagnostic scoring sheet is filled out for almost all of them. This was not the case in the other clinics; some applicants were referred elsewhere or dropped out of the diagnostic process in an earlier stage. As a result, no diagnostic data is available for those patients, implying that the percentages cannot be directly compared as an indication for "strictness." This is a limitation of the study. Another disadvantage of the fact that few diagnostic scoring sheets had been filled out at the time of data analysis in Ghent, Hamburg, and Amsterdam for applicants not fulfilling criteria, is that it compromises the comparison between applicants with and without a diagnosis. We suggest that future (multisite) studies also gather diagnostic information of patients who discontinue the diagnostic process relatively early, as soon as sufficient information is available to do so.

We observed that there were more MtF applicants than FtM applicants. However, a larger percentage of FtM applicants received the GID diagnosis. We found that only one item was gender biased on the basis of our analyses. However, it should be noted that the absence of item bias does not imply that the criteria themselves are equally valid for both sexes [48]. It is conceivable that GID (as any other disorder) appears or expresses itself slightly differently in males and females, and that this is the cause of differences found in GI scores as well as prevalence/incidence as reported in previous studies [5–8,21]. Future studies directed at elucidating this issue are necessary to further facilitate the interpretation of sex differences related to GID.

Conclusion and Recommendations

In the face of our results, we would suggest that it might be helpful for clinicians if the severity and duration of symptoms would be taken into account in the next version of the DSM. The distinction between A and B criteria was not supported by our findings and might have to be reconsidered. Worldwide data-collection that takes severity and duration of the GID symptoms into account would be very helpful in reaching a cross-cultural consensus of how these aspects of symptoms should be weighed in the diagnostic process. Clinicians who participated in our study had trouble interpreting the subcriterion "conviction that he or she has the typical feelings of the other sex," which was expressed in differential item functioning for two items pertaining to this criterion. This might be a reason to remove or rewrite this criterion in the next DSM.

To our knowledge, this is the first study to combine the strength of an international multisite study with the strength of NIRT in order to analyze the generality and utility of a DSM diagnosis. We hope to have convinced the reader of the value of standardized multisite studies, as well as the potential of NIRT, when the focus is on scrutinizing diagnostic criteria. The DSM-5 is currently under development, and whether to enhance the DSM by adding a dimensional adjunct to each of the traditional categorical diagnoses in the DSM is being considered [49]. IRT is likely to play an important role in this enhancement, since it is an excellent method to create dimensional scales and provides a powerful framework for examining the generality of specific symptoms [20,49]. We have shown that these qualities are not limited to parametric IRT and we would recommend that more researchers consider NIRT as an alternative in future studies, especially in studies with small sample sizes and/or with data that show a poor fit to parametric models.

Acknowledgments

This study was supported by the South-Eastern Norway Regional Health Authority and the University of Oslo. The authors would like to thank all clinicians participating in this project for gathering diagnostic data, as well as M. van Duijn, R. Meijer, J. van Bebber, and M. J. Paap for helpful discussions.

Corresponding Author: Muirne C.S. Paap, MS, Department of Neuropsychiatry and Psychosomatic Medicine, Oslo University Hospital, Rikshospitalet Oslo Norway 0027. Tel: +47 23074160; Fax: +47 23074170; E-mail: muirne@nxdomain.nl

Conflict of Interest: None.

Statement of Authorship

Category 1

(a) Conception and Design

Muirne C.S. Paap; Baudewijntje P.C. Kreukels; Peggy T. Cohen-Kettenis; Griet de Cuypere; Hertha Richter-Appelt; Ira R. Haraldsen

(b) Acquisition of Data
Baudewijntje P.C. Kreukels; Peggy T. Cohen-Kettenis; Muirne C.S. Paap; Ira R. Haraldsen; Griet de Cuypere; Hertha Richter-Appelt

(c) Analysis and Interpretation of Data
Muirne C.S. Paap

Category 2

(a) Drafting the Article
Muirne C.S. Paap; Ira R. Haraldsen

(b) Revising It for Intellectual Content
Baudewijntje P.C. Kreukels; Peggy T. Cohen-Kettenis; Griet de Cuypere; Hertha Richter-Appelt

Category 3

(a) Final Approval of the Completed Article
Muirne C.S. Paap; Ira R. Haraldsen; Baudewijntje P.C. Kreukels; Peggy T. Cohen-Kettenis; Griet de Cuypere; Hertha Richter-Appelt

References

1 WHO. The ICD–10 classification of mental and behavioral disorders: Clinical descriptions and diagnostic guidelines. Geneva: World Health Organization; 1992.
2 APA. Diagnostic and statistical manual of mental disorders (4th edn, text revision) (DSM–IV–TR). Washington, DC: American Psychiatric Association; 2000.
3 Weyers S, Elaut E, Sutter PD, Gerris J, T'Sjoen G, Heylens G, Cuypere GD, Verstraelen H. Long-term assessment of the physical, mental, and sexual health among transsexual women. J Sex Med 2009;6:752–60.
4 Fisher AD, Bandini E, Ricca V, Ferruccio N, Corona G, Meriggiola MC, Jannini EA, Manieri C, Ristori J, Forti G, Mannucci E, Maggi M. Dimensional profiles of male to female gender identity disorder: An exploratory research. J Sex Med 2010;7:2487–98.
5 Okabe N, Sato T, Matsumoto Y, Ido Y, Terada S, Kuroda S. Clinical characteristics of patients with gender identity disorder at a Japanese gender identity disorder clinic. Psychiatry Res 2008;157:315–8.
6 Gomez-Gil E, Trilla A, Salamero M, Godas T, Valdes M. Sociodemographic, clinical, and psychiatric characteristics of transsexuals from Spain. Arch Sex Behav 2008; 38:378–92.
7 De Cuypere G, Van Hemelrijck M, Michel A, Carael B, Heylens G, Rubens R, Hoebeke P, Monstrey S. Prevalence and demography of transsexualism in Belgium. Eur Psychiatry 2007;22:137–41.
8 Vujovic S, Popovic S, Sbutega-Milosevic G, Djordjevic M, Gooren L. Transsexualism in Serbia: A twenty-year follow-up study. J Sex Med 2008;6:1018–23.
9 Herman-Jeglinska A, Grabowska A, Dulko S. Masculinity, femininity, and transsexualism. Arch Sex Behav 2002;31:527–34.
10 Sommer IEC, Cohen-Kettenis PT, van Raalten T, vd Veer AJ, Ramsey LE, Gooren LJG, Kahn RS, Ramsey NF. Effects of cross-sex hormones on cerebral activation during language and mental rotation: An fMRI study in transsexuals. Eur Neuropsychopharmacol 2008;18:215–21.
11 Haraldsen IR, Egeland T, Haug E, Finset A, Opjordsmoen S. Cross-sex hormone treatment does not change sex-sensitive cognitive performance in gender identity disorder patients. Psychiatry Res 2005;137:161–74.
12 Cohen-Kettenis PT, Delemarre-van de Waal H, Gooren LJG. The treatment of adolescent transsexuals: Changing insights. J Sex Med 2008;5:1892–97.
13 Sohn M, Bosinski HAG. Gender identity disorders: Diagnostic and surgical aspects (CME). J Sex Med 2007;4:1193–208.
14 Kockott G, Fahrner EM. Male-to-female and female-to-male transsexuals: A comparison. Arch Sex Behav 1988;17:539–46.
15 Smith YL, Van Goozen SH, Kuiper AJ, Cohen-Kettenis PT. Sex reassignment: Outcomes and predictors of treatment for adolescent and adult transsexuals. Psychol Med 2005;35:89–99.
16 Paap MCS, Haraldsen IR. Sex-based differences in answering strategy and the influence of cross-sex hormones. Psychiatry Res 2010;175:266–70.
17 Blanchard R, Clemmensen LH, Steiner BW. Heterosexual and homosexual gender dysphoria. Arch Sex Behav 1987;16:139–52.
18 Schöning S, Engelien A, Bauer C, Kugel H, Kersting A, Roestel C, Zwitserlood P, Pyka M, Dannlowski U, Lehmann W, Heindel W, Arolt V, Konrad C. Neuroimaging differences in spatial cognition between men and male-to-female transsexuals before and during hormone therapy. J Sex Med 2010;7:1858–67.
19 Slabbekoorn D, van Goozen SH, Megens J, Gooren LJ, Cohen-Kettenis PT. Activating effects of cross-sex hormones on cognitive functioning: A study of short-term and long-term hormone effects in transsexuals. Psychoneuroendocrinology 1999;24:423–47.
20 Kraemer HC, Shrout PE, Rubio-Stipec M. Developing the diagnostic and statistical manual V: What will "statistical" mean in DSM-V? Soc Psychiatry Psychiatr Epidemiol 2007;42:259–67.
21 Cohen-Kettenis PT, Gooren LJ. Transsexualism: A review of etiology, diagnosis and treatment. J Psychosom Res 1999;46:315–33.
22 Feske U, Kirisci L, Tarter RE, Pilkonis PA. An application of item response theory to the DSM-III-R criteria for borderline personality disorder. J Personal Disord 2007;21:418–33.
23 Gelhorn H, Hartman C, Sakai J, Stallings M, Young S, Rhee SH, Corley R, Hewitt J, Hopfer C, Crowley T. Toward DSM-V: An item response theory analysis of the diagnostic process for DSM-IV alcohol abuse and dependence in adolescents. J Am Acad Child Adolesc Psychiatry 2008;47:1329–39.
24 Langenbucher JW, Labouvie E, Martin CS, Sanjuan PM, Bavly L, Kirisci L, Chung T. An application of item response theory analysis to alcohol, cannabis, and cocaine criteria in DSM-IV. J Abnorm Psychol 2004;113:72–80.
25 Kan CC, Breteler MH, van der Ven AH, Zitman FG. An evaluation of DSM-III-R and ICD-10 benzodiazepine dependence criteria using Rasch modelling. Addiction 1998;93:349–59.
26 Akechi T, Ietsugu T, Sukigara M, Okamura H, Nakano T, Akizuki N, Okamura M, Shimizu K, Okuyama T, Furukawa TA, Uchitomi Y. Symptom indicator of severity of depression in cancer patients: A comparison of the DSM-IV criteria with alternative diagnostic criteria. Gen Hosp Psychiatry 2009; 31:225–32.
27 Harford TC, Yi H, Faden VB, Chen CM. The dimensionality of DSM-IV alcohol use disorders among adolescent and adult drinkers and symptom patterns by age, gender, and race/ethnicity. Alcohol Clin Exp Res 2009;33:868–78.
28 Compton WM, Saha TD, Conway KP, Grant BF. The role of cannabis use within a dimensional approach to cannabis use disorders. Drug Alcohol Depend 2009;100:221–27.
29 Gelhorn H, Hartman C, Sakai J, Mikulich-Gilbertson S, Stallings M, Young S, Rhee SOO, Corley R, Hewitt J, Hopfer C,

Crowley T. An Item Response Theory Analysis of DSM-IV Conduct Disorder. J Am Acad Child Adolesc Psychiatry 2009;48:42–50.
30 Gillespie NA, Neale MC, Prescott CA, Aggen SH, Kendler KS. Factor and item-response analysis DSM-IV criteria for abuse of and dependence on cannabis, cocaine, hallucinogens, sedatives, stimulants and opioids. Addiction 2007;102:920–30.
31 Agrawal A, Nurnberger JI, Lynskey MT. Item response modeling of DSM-IV mania symptoms in two representative US epidemiological samples. Psychol Med 2010;40:1549–58.
32 Uebelacker LA, Strong D, Weinstock LM, Miller IW. Use of item response theory to understand differential functioning of DSM-IV major depression symptoms by race, ethnicity and gender. Psychol Med 2009;39:591–601.
33 Jane JS, Oltmanns TF, South SC, Turkheimer E. Gender bias in diagnostic criteria for personality disorders: An item response theory analysis. J Abnorm Psychol 2007;116: 166–75.
34 Weinstock LM, Strong D, Uebelacker LA, Miller IW. Differential item functioning of DSM-IV depressive symptoms in individuals with a history of mania versus those without: An item response theory analysis. Bipolar Disorders 2009;11:289–97.
35 Kreukels BPC, Haraldsen IR, De Cuypere G, Richter-Appelt H, Gijs L, Cohen Kettenis PT. A European Network for the Investigation of Gender Incongruence: The ENIGI initiative. Eur Psychiatry. 2010 Jul 8 [Epub ahead of print] doi:10.1016/j.eurpsy.2010.04.009.
36 Meyer-Bahlburg H. From mental disorder to iatrogenic hypogonadism: Dilemmas in conceptualizing gender identity variants as psychiatric conditions. Arch Sex Behav 2010; 39:461–76.
37 Mokken RJ. A theory and procedure of scale analysis. the Hague: Mouton; 1971.
38 Sijtsma K, Molenaar IW. Introduction to nonparametric item response theory. Thousand Oaks: Sage Publications; 2002.
39 Molenaar IW, Sijtsma K. MSP5 for Windows. Groningen, the Netherlands: iecProGAMMA; 2000.
40 Sijtsma K, Emons WH, Bouwmeester S, Nyklicek I, Roorda LD. Nonparametric IRT analysis of Quality-of-Life Scales and its application to the World Health Organization Quality-of-Life Scale (WHOQOL-Bref). Qual Life Res 2008; 17:275–90.
41 Wismeijer AA, Sijtsma K, van Assen MA, Vingerhoets AJ. A comparative study of the dimensionality of the self-concealment scale using principal components analysis and Mokken scale analysis. J Pers Assess 2008;90:323–34.
42 Meijer RR, Baneke JJ. Analyzing psychopathology items: A case for nonparametric item response theory modeling. Psychological Methods 2004;9:354–68.
43 Mokken RJ. Nonparametric models for dichotomous responses. In: van der Linden WJ, Hambleton RK, eds. Handbook of modern item response theory. New York: Springer; 1997;351–67.
44 Ramsay JO. Testgraf. A program for the analysis of multiple choice test and questionnaire data. Montreal, Canada: Department of Psychology, McGill University; 2000.
45 Molenaar IW. Nonparametric models for polytomous responses. In: van der Linden WJ, Hambleton RK, eds. Handbook of modern item response theory. New York: Springer; 1997;369–80.
46 Sijtsma K, Meijer RR. A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. Appl Psych Meas 1992;16:149–57.
47 SPSS. SPSS for Windows, Rel. 16.0.1. Chicago: SPSS Inc.; 2007.
48 Hartung CM, Widiger TA. Gender differences in the diagnosis of mental disorders: Conclusions and controversies of the DSM-IV. Psychol Bull 1998;123:260–78.
49 Kraemer HC. DSM categories and dimensions in clinical and research contexts. Int J Methods Psychiatr Res 2007;16:S8–S15.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Inclusion criteria and diagnostic ratings.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.