An Item Response Theory Analysis of the Forensic Restrictiveness Questionnaire (FRQ)

Dr. Jack Tomlin (https://orcid.org/0000-0002-7610-7918)[1], Prof. Peter Bartlett[2], Dr. Vivek Furtado[3], Dr. Vincent Egan (https://orcid.org/0000-0003-0878-2556)[4], Prof. Dr Birgit Völlm, PhD (https://orcid.org/0000-0003-4571-3410)[1]

[1] Department of Forensic Psychiatry, University of Rostock, Rostock, Germany

[2] School of Law and Institute of Mental Health, University of Nottingham, Nottingham, UK

[3] Mental Health and Wellbeing, Warwick Medical School, University of Warwick, Coventry, UK

[4] Centre for Family and Forensic Psychology, University of Nottingham, Nottingham, UK

Corresponding Author:
Jack Tomlin
Klinik für Forensische Psychiatrie, Gehlsheimer Straße 20, 18147, Rostock, Germany
Email: jack.tomlin@med.uni-rostock.de

Abstract

Forensic psychiatric care settings are intended to be more therapeutic than penal settings. They attempt to be more homely, recovery-oriented, person-centered, and less overtly punitive. However, forensic inpatient settings are still highly secure, risk averse, and diminutive of patient autonomy. Accordingly, a body of research is investigating how patients experience their care and how these experiences are associated with treatment outcomes. The self-report Forensic Restrictiveness Questionnaire (FRQ) is a 15-item questionnaire of patients' perceptions of the restrictions upon their autonomy. There has been interest in validating the FRQ in several countries.

Despite the promising preliminary empirical support for the FRQ, its psychometric properties are not well understood. In this paper we draw on Item-Response Theory (IRT) to investigate the properties of individual FRQ items to identify candidate items for alteration, removal or retention to assist researchers validating the FRQ in new contexts. The results suggest the FRQ is more sensitive to measuring the perceptions of patients that have above average amounts of restrictiveness. Measurement error rises sharply for the approximately 5% highest scoring respondents but is low for the majority of individuals. Users are likely to respond in a dichotomised manner and not use the 'Not Sure' option. The response category 'Not Applicable' should be removed from a revised FRQ.

1. INTRODUCTION

An emerging body of research is exploring to what extent and in what ways forensic mental health patients are restricted in their care (Tomlin, Bartlett, & Völlm, 2018; Tomlin, Bartlett, Völlm, Furtado, & Egan, 2020). The genesis for this work lies in the determination made in many jurisdictions that psychiatric care should be provided in the *least restrictive setting* appropriate (Appelbaum 1999). This maxim reflects a societal emphasis on patients' rights and the recognition of individuals living with mental illness as empowered, capable of living alongside mental ill-health in the community (Anthony, 1993). It further reflects developments in pharmaceuticals facilitating the self-management of symptoms, scientific research identifying the harms resulting from long-term institutionalization, and contemporary approaches to mental health care and offender rehabilitation (Chow & Priebe, 2013; Goffman, 1961; Pouncey & Lukens, 2013).

A sub-section of this work has looked into the prevalence and consequence of coercion and coercive measures. Coercive measures, typically defined as restraint, seclusion, forced medication or involuntary admission, are used frequently in forensic services (Hui et al. 2013). More broadly understood, coercion has been defined as the use of threats, pressure or denying patients choice (Szmukler, 2015). Coercion and coercive measures might be enacted in the best interests of the targeted person or for the protection of others (Soininen, Kontio, Joffe, & Putkonen, 2016). Coercion might further be used for discipline, order or preventing escape (Albrecht, 2016). These definitions typically share the requirement of an action; that the coercion is intended by others and restricts the autonomy of the coerced (Newton-Howes & Banks, 2014).

Coercive measures have predominantly been linked to negative outcomes, however. Coercion can lead patients to feel there was a violation of their human rights, disrespected, not heard by clinicians, and dehumanized (Newton-Howes & Banks, 2014). Patients that have experienced perceived coercion are less likely to consider staff as helpful, hindering the development of positive therapeutic relationships (Cope & Encandela, 1998). Kontio et al. (2012) found coercive measures undermined satisfaction in care, treatment adherence, and patient autonomy. The use of coercive measures is contentious given these outcomes and the dearth of controlled trials investigating their clinical efficacy (Elcock & Lewis, 2016).

A related body of research has investigated forensic patients' subjective experiences of restrictions upon their autonomy. The rationale for this is that patients are likely to experience care as restrictive in different ways and that is it important to explore how this affects individuals' recovery. Sustere and Tarpey (2019) interviewed patients in an English medium security hospital on the use of least restrictive practices. Patients reported feeling that the setting limited their opportunities for social interaction, which made them lonely; they felt unable to take control of aspects of their care including contributing to their risk assessments. Hui (2017) interviewed patients in an English high secure setting. Patients in her study described restrictive practices as including close confinement with others, lacking a private space, and having few personal belongings. They described these as humiliating, fear-provoking, degrading, inhumane, anxiety-provoking, and detrimental to dignity, suggesting the setting engendered dependence.

We investigated how patients in low, medium and high security English hospitals described experiences of restrictiveness (Tomlin, Egan, Bartlett, & Völlm, 2019). Patients attributed restrictive elements of care to the level of risk they felt staff thought they posed, the amount of resources the setting had, and whether forensic care was oriented towards treatment or punishment. Salient restrictions were grouped into four main themes: limitations that affected daily life on the ward; the (indefinite) length of stay; difficulties navigating old and developing new relationships; and the ways in which their identity was reshaped by life in secure care. Interviewees described feeling institutionalized, therapized, bored, and frustrated.

1.01 The Forensic Restrictiveness Questionnaire

Without valid measurement instruments, it is not possible to infer with much confidence the empirical association between experiences of restrictiveness and recovery outcomes. Accordingly, we developed the self-report Forensic Restrictiveness Questionnaire (FRQ; Tomlin, Völlm et al (2019)). The FRQ is novel as it provides the first self-report measure of patients' subjective experiences of restrictiveness using questions derived from qualitative interviews (Tomlin, Egan et al. 2019). The FRQ captures a unidimensional construct with 15, 5-point Likert-scaled statements. Respondents are asked to what extent they agree or disagree with statements such as the following: *'I am treated like a human being here'*; *'I can express my feelings here enough'*; and *'I can choose what I want to do each day'*. All but one item is positively worded and the questionnaire takes approximately 5 minutes to compete.

The FRQ demonstrated sound psychometric properties (Tomlin, Völlm, et al., 2019). The original pilot FRQ with 58 items was completed by 235 patients in low, medium and high secure settings across England. Exploratory Factor Analysis using Principle Axis Factoring and Oblique, PROMAX rotation suggested a single latent factor was most appropriate for explaining variance. Internal consistency was high (Cronbach's a = .93). There was evidence for convergent validity as the FRQ was negatively associated with both quality of life as measured by the Forensic Inpatient Quality of Life Questionnaire – Short Version (FQL-SV; Schel, Bouman, Vorstenbosch, & Bulten (2017)) ($\rho = -0.61$, p < .001, n = 229) and ward atmosphere, measured with EssenCES (Schalast, Redies, Collins, Stacey, & Howells (2008)) ($\rho = -0.72$, p < .001, n = 229). Qualitative feedback from participants was reviewed and incorporated into the final 15-item FRQ. There are plans to adapt the FRQ for use in Italy, Poland, Germany, Portugal, Turkey, Norway and Canada.

2. AIMS AND RATIONALE

Despite the promising preliminary empirical support for the utility of the FRQ, its psychometric properties are not well understood as there are no studies published beyond those relating to its initial validation. To better our knowledge of the FRQ, the present paper investigates these properties further. The data described in this paper are taken from the pilot study but provide helpful information for researchers scrutinizing the applicability, reliability and validity of the FRQ in their local context.

This paper describes the utility of each item included in the FRQ. In doing so, we draw on Item-Response Theory (IRT), specifically Graded Response Modelling (GRM), to investigate the probabilities of participant responses across item response categories given a participants' restrictiveness score, the difficulty and discrimination parameters of each item, and the amount of information contained in each item. The aim was to identify candidate items for alteration, removal or retention. Researchers looking to adapt the FRQ can make better-informed decisions about whether to remove or keep items.

3. METHODS

3.01 Item Response Theory

IRT models calculate the probabilities for how individuals will respond to a particular survey item given their score on a latent trait, ability or outcome (Cooper, 2018). IRT models calculate a difficulty parameter for each item based on how people in a sample responded to it and all other items. If the majority of respondents correctly answer an item but incorrectly answer another item, IRT modelling assumes the latter is more difficult or requires a higher amount of a latent concept. Thus, if someone correctly answers a very difficult item, we can predict they may answer most of the easier questions correctly too – this is an application of Guttmann scaling (Cooper, 2018). The greater the number of difficult questions someone answers correctly, the more validly we can predict that they possess a larger amount of a trait the items aim to measure (e.g. mathematical ability, perceptions of restrictiveness). The difficulty level of an item is called its 'difficulty parameter'. The probability of correctly answering an item is the product of an individual’s trait score and the item’s difficulty parameter (Thorpe & Favia, 2012).

The most basic IRT model is the one-parameter Rasch model (1PL; Reeve and Fayers (2005)). This assumes that items are dichotomous and can be answered correctly or incorrectly. The ‘one-parameter’ refers to the difficulty parameter, or ‘*b*’ (Reeve & Fayers, 2005). The trait being measured is referred to as Theta, or ‘θ’, depicted on graphs as an x-axis continuum marked by standard deviations (see Figure 1). An extension to this model, the 2PL Rasch model incorporates a second important parameter: the discrimination parameter, or ‘*a*’ (Reeve & Fayers, 2005). This parameter tells us to what extent an item can distinguish between respondents at different levels of Theta. If an item has low discrimination, then a large increase in the measured trait is needed to see a meaningful increase in the probability of a correct response. These parameters are depicted graphically by Item Characteristic Curves (ICCs). Graded Response Modelling is a derivation of 2PL Rasch modelling that is suitable for polytomous data (i.e. 5-point Likert scales; Samejima, 1969). GRM uses similar concepts but applies different mathematic formulae to account for the greater range of responses.

Instead of providing ICCs, GRM produces Boundary Characteristic Curves (BCCs). These illustrate the discrimination and difficulty parameters for each item response category. Each BCC represents the probability of a respondent selecting all response categories below and including a certain category versus the probability of selecting any response category that is higher (Gomez & Fisher, 2005). For example, the probability of selecting 1= ‘Strongly Disagree’ *versus* 2= ‘Agree’ and 3= ‘Not Sure’ and 4= ‘Agree’ etc. is depicted by a single BCC. Subsequently, the probability of selecting 1 and 2 *versus* 3 and 4 etc. is depicted by a second BCC. The difficulty parameter for each BCC is where the probability of choosing either option in these dichotomies is 50%; this is depicted by the dotted horizontal lines in Figure 1. The discrimination parameter for each category is depicted in the slope of each line – the more sloped, the lower the discrimination value.

GRM provides Category Characteristic Curves (CCCs), as shown in Figure 2. CCCs are more intuitive than BCCs as they depict the probabilities of individuals answering each response category given their Theta score. CCCs are helpful for assessing item utility as they can identify response categories that are selected by individuals many standard deviations from the Theta mean or for which there is an extremely low probability of ever being selected by a respondent regardless of their Theta value.

GRM also provides Item and Test Information Functions (IIFs and TIFs), as presented in

Figures 3 and 4. IIFs indicate the Theta values for which an item is best at measuring respondents' trait levels; TIFs provide this for the test as a whole (Cooper, 2018; Thorpe & Favia, 2012). The shape of an item's information function is the product of its discrimination and difficulty parameters. Higher information functions suggest an item is more precise. In fact, the inverse of an item or measure's information function is an indicator of measurement error. A good scale has items with high information functions that spread across the Theta continuum. Two items with identical IIFs are candidates for removal as one may be redundant.

GRM holds several assumptions: scale unidimensionality, monotonically increasing item characteristic curves (or BCCs), and item local independence (Reeve & Fayers, 2005; StataCorp, 2017). Unidimensionality implies that the questionnaire items measure a single latent trait. Local independence provides that all items are independent from each other and that a response to one item does not influence the response to another. Monotonicity implies that as the value of the latent trait increases, so too does the probability of a dominant answer on the item response, in a non-decreasing manner (Stochl, Jones, & Croudace, 2012). Finally, sample sizes for GRM need to be quite large. Thorpe and Favia (2012) report mixed findings in the literature but suggest that N=250 is needed for polytomous modelling but N=500 is recommended for more accurate parameter estimates.

In summary, GRM is able to provide researchers information that helps inform the removal or retention of questionnaire items, including difficulty and discrimination parameters, the probabilities of responses across categories given individuals' trait score, and both item and test information functions.

3.02 Participant Recruitment

IRT models do not require random samples; groups with low Theta scores will produce the same B/ICCs as groups with high Theta scores (Thorpe & Favia, 2012). The sampling frame for this project comprised forensic in-patients over the age of 18 in England. Recruitment took a stratified, convenience approach (Lynn, 2016). Patients in different secure wards and hospitals were approached to capture a range of diagnoses, levels of security, and treatment progress. A member of the research team presented the aims of the project to patients and staff at ward meetings and arranged to meet with patients at a later date. Potential participants were given information sheets and at least 24 hours to reconsider their involvement. All participants signed consent forms. The exclusion criteria were: insufficient understanding of English or without access to a translator and no capacity to consent and participate. A primary diagnosis of learning disability or being under the age of 18 were exclusion criteria. The National Health Service

Clinical Research Network helped recruit participants.

3.03 Data Collection

The pilot 58-item FRQ was completed by 235 in-patients in low, medium and high security forensic hospitals in England. Participants were handed the pilot FRQ to complete with the assistance of a member of the research team. Demographic, clinical and legal data were collected by the researchers from patient notes. The pilot FRQ was validated according to Classical Test Theory principles, and data from the resulting 15-item validated FRQ were used in the present study (Tomlin, Völlm, et al., 2019). The majority of respondents were male (96%), white-British (70%), and diagnosed with a psychotic disorder under ICD-10 F.2 (60%). Mean age was 39.8 years; median length of stay was 19 months. The most commonly committed offence-type was against the person (37%), and the largest group of patients was given a hospital order for treatment with legal restrictions by a criminal court (43%). Full details are reported in Tomlin, Völlm et al. (2019).

3.04 Data Analysis

Missing data were assessed with SPSS v. 24. Missing data accounted for 0.26% of the pilot FRQ data. According to Little’s test of missing completely at random (MCAR), the missing data were missing at random and thus suitable for multiple imputation ($\chi^2(639) = 530.860$, $p = 0.999$). Accordingly, missing data were inputted via SPSS’s multiple imputation function (Tabachnick & Fidell, 2013). Items were coded as follows: ‘0’ = Not Applicable, ‘1’ = Strongly Disagree; ‘2’ = Disagree a Little; ‘3’ = Not Sure; ‘4’ = Agree a Little; ‘5’ = Strongly Agree. All items were coded so that a higher score indicates a greater amount of restrictiveness. STATA v. 15 was used for the analysis. STATA’s Item-Response Theory, Graded Response Model function was used with the command ‘irt grm [vars]’ (StataCorp, 2017). Results were significant at $p<0.05$.

In the present study, the following traits were investigated for each FRQ item:

Difficulty parameters. These were interpreted visually via BCCs and quantitatively. Item response categories with difficulty parameters lower than -3 or greater than +3 standard deviations from the Theta mean were considered likely redundant and were candidate items for alteration or removal. This was for the simple reason that they would only distinguish between a very low number of respondents scoring on the extreme ends of the Theta continuum, or 0.6% of responses, and add little information to the FRQ.

Discrimination parameter. These were interpreted visually via BCCs and quantitatively. Discrimination parameter values from 0.10 to .34, and then including and above the thresholds 0.35, 0.65, 1.35 and 1.70 were considered very low, low, moderate, high and very high respectively (Thorpe & Favia, 2012). Items with discrimination parameters under 0.64 were candidate items for alteration or removal.

Responses across item categories. These were interpreted visually via CCCs. Item response categories that were always less likely to be selected by respondents than any other response category across the Theta continuum were considered likely redundant and were candidate items for alteration or removal.

Item and test information functions. These were interpreted visually via IIFs and TIFs. Items with item information functions that were visually almost identical to a second item were candidate items for alteration or removal.

Recommendations on the alteration or removal of items were made by assessing the results across all four traits. In other words, having a low discrimination parameter was not a sufficient ground for item alteration or removal if the item was of strong theoretical relevance to the FRQ and it scored well on the other three traits.

3.05 Ethical Approval

Ethical approval for the study was granted by the Leicestershire South Research Ethics Committee and the Health Research Authority of the NHS. Study reference code: 17/EM/0159.

4. RESULTS

4.01 Assumptions

The FRQ is a unidimensional scale (Tomlin, Völlm, et al., 2019). None of the items affect an individual’s response to another item and accordingly the items demonstrate local independence. This is shown in the BCCs for all items in Figure 1, where the probability of a dominant response to an item increased with Theta. The three assumptions of GRM modeling were met. The sample size of 235 was just short of the N=250 suggested in Thorpe and Favia (2012). The sample size should be considered a weakness in the present study. However, given the unidimensionality and high internal consistency of the FRQ, it is was considered justified to proceed with the analysis and interpret the results with this in mind.

4.02 Discrimination Parameters

Table 1 displays the discrimination parameters (*a*) for each item. The lowest discrimination parameter is for FRQ13 (1.35) and the highest is for FRQ1 (2.46). These are high to very high parameter values. The slopes of the BCCs for each item are shown in Figure 1. We can observe that the BCCs are moderately steep, supporting the parameter values and suggesting that the FRQ items and response categories are good at differentiating among respondents at different levels of the Theta continuum.

Table 1 about here

4.03 Difficulty Parameters

The difficulty parameters are given in Table 1. These are also depicted graphically as BCCs in Figure 1. Two key observations can be made from observing the item BCCs. First, 75% of the difficulty parameters (threshold parameters) were located between -1 and +2 standard deviations from the mean Theta score. This is a fairly even spread around the Theta mean but the BCCs indicate that many response categories have an above average difficulty parameter.

Second, in the items with five BCCs, the difficulty parameters of the leftmost BCCs are all under -2.5 standard deviations below the Theta mean. This reflects the inclusion of the ‘Not Applicable’ response category. This is informative as it suggests that in items including the ‘Not Applicable’ category, respondents scoring -3 standard deviations below the Theta (restrictiveness) mean and greater have a nearly 100% probability of responding with a ‘higher’ response category i.e. ‘Strongly Agree’ etc. Therefore, it is only people with extremely low amounts of Theta (restrictiveness) for which the ‘Not Applicable’ item is informative.

4.04 Responses across Item Categories

Figure 2 displays the Category Characteristic Curves (CCCs) for each item. Three observations can be made. First, for the items where at least one individual responded with ‘Not Applicable’, the 50% probability point at which a respondent is likely to select this category is positioned around -3 standard deviations below the Theta mean as described in the BCCs. This again suggests that only a few respondents experiencing very little restrictiveness are likely to choose this category.

The second peak suggests that the FRQ items are most informative about individuals with an above-average restrictiveness score and less informative for people with Theta scores less than -1 standard deviation below the mean. This supports the grouping of BCCs in Figure 1. These findings are further supported by the TIF shown in Figure 4, which indicates that measurement error rises sharply for the approximately 5% highest scoring respondents. Measurement error is slightly raised for items measuring around -2 standard deviations below the mean but is low for a large part of the Theta continuum.

Figure 1 Boundary Characteristic Curves about here

Second, the ‘Strongly Agree’ and ‘Strongly Disagree’ categories are more likely to be selected than other categories. This suggests that respondents are more likely to respond in a dichotomized manner despite the polytomous response structure of each item. Third, in all items there is no Theta value for which a respondent is more likely to choose the ‘Not Sure’ option than any other. All ‘Not Sure’ CCCs fall below all other CCCs. This suggests that the ‘Not Sure’ category does not contribute much useful information or measure respondents’ experiences helpfully.

Figure 2 Category Characteristic Curves about here

4.05 Item and Test Information Functions

The IIFs for all items are displayed in Figure 3. Two overall observations can be made. First, there are two peaks. This suggests that the FRQ items provide most information for people scoring around -3 standard deviations below Theta mean, representing an extremely low 0.23% of respondents, and for people scoring just under the mean until +2 standard deviations above it. The first peak is likely explained by the items that include a ‘Not Applicable’ category; this category on these items helps identify those individuals that really do not feel restricted.

The second observation is that some items provide less information than others but follow similar IIF patterns. For example, items 3, 10 and 13 appear to contain a similar range of information to each other and are less informative at the above-average Theta range than items 4 and 1 for instance. These items might be candidates for removal if administrators are eager to shorten the FRQ, but given its current length and brevity of completion, we do not believe shortening the FRQ should be a goal. Further studies should confirm the utility of these items after the ‘Not Sure’ and ‘Not Applicable’ response categories are removed.

Figure 3 Item Information Functions about here

Figure 4 Test Information Function about here

5. DISCUSSION

5.01 Implications for Revisions to the FRQ

This study investigated the psychometric properties of the Forensic Restrictiveness Questionnaire from an Item Response Theory perspective. This was undertaken with the aim of identifying items that might benefit from revision and removal. The findings complement a previous validation study that used Classical Test Theory methods. Graded Response Modelling was used on a sample of 235 forensic patients in England. The results suggest that some changes can be made to the FRQ to improve is usefulness and psychometric properties.

The FRQ items demonstrated strong discrimination parameters. Values ranged from 1.35 to 2.46. According to the standards described by Thorpe and Favia (2012) values greater than 1.35 indicate high discrimination, with values over 1.70 considered very high. This suggests the FRQ items are good at distinguishing between people with levels of restrictiveness lesser or greater than the difficulty of each item. The difficulty parameters of the FRQ items were mostly positioned between -1 and +2 standard deviations from the Theta mean. Specifically, 75% of all response categories fell within this range; 85% of responses when excluding the 'Not Applicable' category. A revised version of the FRQ might therefore benefit from including items that measure lower levels of restrictiveness, perhaps by asking about elements of care that are less frequently described by patients as restrictive. The difficulty parameters suggest the FRQ is more sensitive to measuring respondents with above average perceptions of restrictiveness.

The BCCs depicting the 50% probability that respondents selected the 'Not Applicable' category *versus* all other categories were skewed towards respondents with very low Theta scores. In fact, all difficulty parameters for this category fell below -2.5 standard deviations and 75% under -3 standard deviations. This finding suggests that for the vast majority of patients 'Not Applicable' adds little useful information. The CCCs offered further insight into how the response categories to the FRQ might be revised. Again, these make it apparent that the 'Not Applicable' category offers little helpful information given only respondents with very low Theta scores have a 50% likelihood of selecting it. The 'Not Applicable' category should be removed from the FRQ.

The results also suggest that respondents are more likely to give dichotomised responses than make use of the full Likert scale options. Respondents are still likely to respond with 'Agree' and 'Disagree' and thus these categories are helpful to keep. However, given that respondents were never more likely to select the 'Not Sure' category than any other, this option adds little useful information to the FRQ and might be removed. This would reflect the absence of 'Not Sure' categories in similar questionnaires used in forensic settings (e.g. EssenCES: Schalast et al. (2008); FQL-SV: Schel et al. (2017)) and guidance on survey development that suggests any mid-point on a Likert scale should not reflect an inability to answer but a middle amount of the latent trait (Streiner & Norman, 2008).

Finally, the item and test information functions support the difficulty parameter findings. The FRQ is most competent at measuring respondents between -1 and +2 standard deviations from the Theta mean. Items on which respondents selected ‘Not Applicable’ were somewhat adept at identifying the approximately 0.23 per cent of patients likely to score around -3 standard deviations below the Theta mean. However, given the small clinical or research utility that derives from this, it is not recommended this response category be kept.

5.02 Limitations

This study has two limitations. First, the sample size falls slightly under the number prescribed for polytomous IRT modelling (Thorpe & Favia, 2012). Interpretations of the parameter estimates should be made with this in mind. However, the sample size achieved in the study is comparable to or larger than other studies piloting questionnaires in forensic settings and is also a strength of this study. Second, the data used in this paper are the same as those collected in the pilot study. Follow-up research should administer the FRQ with a new sample and re-run the IRT analysis to confirm or refute the conclusions of this paper.

6. Conclusion

The FRQ is more sensitive to measuring the perceptions of patients that have above average amounts of the restrictiveness latent trait. Measurement error rises sharply for the approximately 5% highest scoring respondents but is low for the majority of individuals’ restrictiveness scores. Users are likely to respond in a dichotomised manner and not use the ‘Not Sure’ option. They do not forego the full Likert scale range, however, and at least four response categories should be kept. The ‘Not Applicable’ response category is only useful for identifying people scoring -3 standard deviations below the mean Theta (restrictiveness) score. Accordingly, the response category ‘Not Applicable’ should be removed from a revised FRQ; future studies should reassess the utility of the ‘Not Sure’ category to see if the results of this study are reproduced.

1. References

Albrecht, H. J. (2016). Legal aspects of the use of coercive measures in psychiatry. *The Use of Coercive Measures in Forensic Psychiatric Care: Legal, Ethical and Practical Challenges*, 31–48. https://doi.org/10.1007/978-3-319-26748-7_3

Anthony, W. A. (1993). Recovery from mental illness: The guiding vision of the mental health service system in the 1990s. *Psychosocial Rehabilitation Journal*, *16*(4), 11. https://doi.org/10.1037/h0095655

Appelbaum, P. S. (1999). Law & Psychiatry: Least restrictive alternative revisited: Olmstead's uncertain mandate for community-based care. *Law & Psychiatry*, *50*(10), 1271-1272,1280. https://doi.org/https://doi.org/10.1176/ps.50.10.1271

Chow, W. S., & Priebe, S. (2013). Understanding psychiatric institutionalization: a conceptual review. *BMC Psychiatry*, *13*(1), 169. https://doi.org/10.1186/1471-244x-13-169

Cooper, C. (2018). *Psychological testing: theory and practice*. Retrieved from https://books.google.co.uk/books?hl=en&lr=&id=XmNwDwAAQBAJ&oi=fnd&pg=PT19&dq=Psychological+testing:+Theory+and+Practice+Colin+Cooper&ots=1ITieM01an&sig=6kBUKqSgG3ktIdIGm8Wzkkyh8zM#v=onepage&q=Psychological testing%3A Theory and Practice Colin Cooper&f=fal

Cope, K. A., & Encandela, J. A. (1998). Organizational features affecting the use of coercion in the administration of psychiatric care. *Adm Policy Ment Health*, *25*(3), 309–319.

Elcock, S., & Lewis, J. (2016). Mechanical restraint: Legal, ethical and clinical issues. *The Use of Coercive Measures in Forensic Psychiatric Care: Legal, Ethical and Practical Challenges.*, 315–331. https://doi.org/http://dx.doi.org/10.1007/978-3-319-26748-7_17

Goffman, E. (1961). Asylums : essays on the social situation of mental patients and other inmates, 386.

Gomez, R., & Fisher, J. W. (2005). Item response theory analysis of the spiritual well-being questionnaire. *Personality and Individual Differences*, *38*(5), 1107–1121. https://doi.org/10.1016/j.paid.2004.07.009

Hui, A. (2017). *Least restrictive practices: an evaluation of patient experiences*. University of Nottingham. Retrieved from http://eprints.nottingham.ac.uk/48816/

Hui, A., Middleton, H., & Vollm, B. (2013). Coercive measures in forensic settings: Findings from the literature. *The International Journal of Forensic Mental Health*, *12*(1), 53–67. https://doi.org/http://dx.doi.org/10.1080/14999013.2012.740649

Kontio, R., Joffe, G., Putkonen, H., Kuosmanen, L., Hane, K., Holi, M., & Valimaki, M. (2012). Seclusion and restraint in psychiatry: patients' experiences and practical suggestions on how to improve practices and use alternatives. *Perspect Psychiatr Care*, *48*(1), 16–24. https://doi.org/10.1111/j.1744-6163.2010.00301.x

Lynn, P. (2016). Principles of Sampling. In T. Greenfield & S. Greener (Eds.), *Research methods for* *postgraduates*. Retrieved from https://ebookcentral.proquest.com/lib/nottingham/reader.action?docID=4644084&ppg=268

Newton-Howes, G., & Banks, D. (2014). The subjective experience of community treatment orders: patients' views and clinical correlations. *The International Journal of Social Psychiatry*, *60*(5), 474–481. https://doi.org/http://dx.doi.org/10.1177/0020764013498870

Pouncey, C. L., & Lukens, J. M. (2013). Madness versus badness: The ethical tension between the recovery movement and forensic psychiatry. In *Applied ethics in mental health care: An interdisciplinary reader* (pp. 237–253). Cambridge, MA: MIT Press; US. https://doi.org/http://dx.doi.org/10.7551/mitpress/9780262019682.003.0017

Reeve, B., & Fayers, P. (2005). Applying item response theory modelling for evaluating questionnaire item and scale properties. *Assessing Quality of Life in Clinical Trial: Methods and Practice*, 55–74. https://doi.org/10.1007/s11136-007-9198-0

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph Supplement No. 17*.

Schalast, N., Redies, M., Collins, M., Stacey, J., & Howells, K. (2008). EssenCES, a short questionnaire for assessing the social climate of forensic psychiatric wards. *Criminal Behaviour and Mental Health*, *18*(1), 49–58. https://doi.org/10.1002/cbm.677

Schel, S. H. H., Bouman, Y. H. A., Vorstenbosch, E. C. W., & Bulten, B. H. (2017). Development of the forensic inpatient quality of life questionnaire: short version (FQL-SV). *Quality of Life Research*, *26*(5), 1153–1161. https://doi.org/10.1007/s11136-016-1461-9

Soininen, P., Kontio, R., Joffe, G., & Putkonen, H. (2016). Patient experience of coercive measures. *The Use of Coercive Measures in Forensic Psychiatric Care: Legal, Ethical and Practical Challenges.*, 255–270. https://doi.org/http://dx.doi.org/10.1007/978-3-319-26748-7_14

StataCorp. (2017). *STATA Item Response Theory Reference Manual, Release 15*.

Stochl, J., Jones, P. B., & Croudace, T. J. (2012). Mokken scale analysis of mental health and well-being questionnaire item responses: a non-parametric IRT method in empirical research for applied health researchers. *BMC Medical Research Methodology*, *12*(1), 74.

Streiner, D. L., & Norman, G. R. (2008). *Health Measurement Scales*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199231881.001.0001

Sustere, E., & Tarpey, E. (2019, January 20). Least restrictive practice: its role in patient independence and recovery. *Journal of Forensic Psychiatry and Psychology*, pp. 1–16. https://doi.org/10.1080/14789949.2019.1566489

Szmukler, G. (2015, October). Compulsion and "coercion" in mental health care. *World Psychiatry*, *14*(3), 259–261. https://doi.org/10.1002/wps.20264

Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics*. *Harper Collins* (6th ed.). New Jersey: Pearson. https://doi.org/10.1037/022267

Thorpe, G. L., & Favia, A. (2012). Data Analysis Using Item Response Theory Methodology: An Introduction to Selected Programs and Applications. *Psychology*, *20*(July), 1–34. https://doi.org/10.1039/9781849734929-00016

Tomlin, J., Bartlett, P., & Völlm, B. (2018). Experiences of restrictiveness in forensic psychiatric care: Systematic review and concept analysis. *International Journal of Law and Psychiatry*, *57*, 31–41. https://doi.org/http://dx.doi.org/10.1016/j.ijlp.2017.12.006

Tomlin, J., Bartlett, P., Völlm, B., Furtado, V., & Egan, V. (2020). Perceptions of Restrictiveness in Forensic Mental Health: Do Demographic, Clinical, and Legal Characteristics Matter? *International Journal of Offender Therapy and Comparative Criminology*. https://doi.org/10.1177/0306624X20902050

Tomlin, J., Egan, V., Bartlett, P., & Völlm, B. (2019). What Do Patients Find Restrictive About Forensic Mental Health Services? A Qualitative Study. *International Journal of Forensic Mental Health*, 1–13. https://doi.org/10.1080/14999013.2019.1623955

Tomlin, J., Völlm, B., Furtado, V., Egan, V., & Bartlett, P. (2019). The Forensic Restrictiveness Questionnaire: Development, Validation, and Revision. *Frontiers in Psychiatry*, *10*. https://doi.org/10.3389/fpsyt.2019.00805