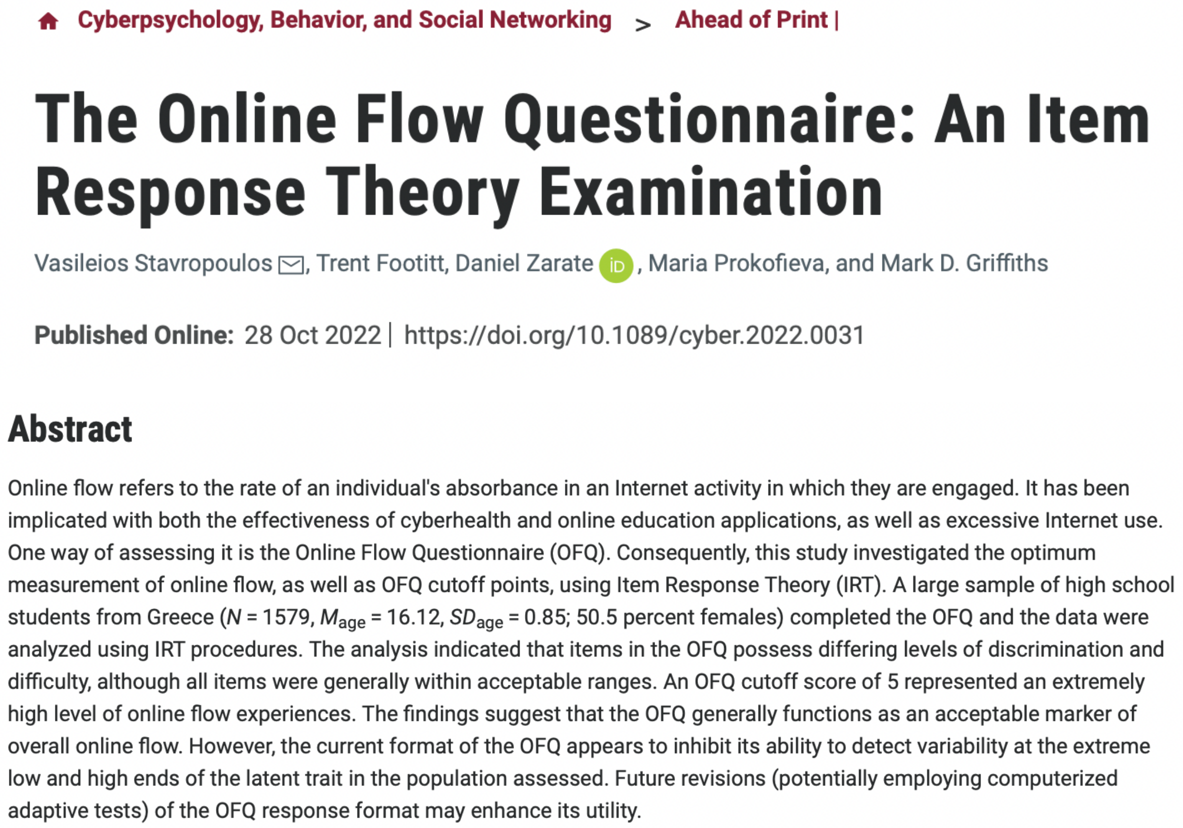


Cyberpsychology, Behavior, and Social Networking > Ahead of Print |

The Online Flow Questionnaire: An Item Response Theory Examination

Vasileios Stavropoulos, Trent Footitt, Daniel Zarate, Maria Prokofieva, and Mark D. Griffiths

Published Online: 28 Oct 2022 | https://doi.org/10.1089/cyber.2022.0031

Abstract

Online flow refers to the rate of an individual's absorbance in an Internet activity in which they are engaged. It has been implicated with both the effectiveness of cyberhealth and online education applications, as well as excessive Internet use. One way of assessing it is the Online Flow Questionnaire (OFQ). Consequently, this study investigated the optimum measurement of online flow, as well as OFQ cutoff points, using Item Response Theory (IRT). A large sample of high school students from Greece (N = 1579, M_{age} = 16.12, SD_{age} = 0.85; 50.5 percent females) completed the OFQ and the data were analyzed using IRT procedures. The analysis indicated that items in the OFQ possess differing levels of discrimination and difficulty, although all items were generally within acceptable ranges. An OFQ cutoff score of 5 represented an extremely high level of online flow experiences. The findings suggest that the OFQ generally functions as an acceptable marker of overall online flow. However, the current format of the OFQ appears to inhibit its ability to detect variability at the extreme low and high ends of the latent trait in the population assessed. Future revisions (potentially employing computerized adaptive tests) of the OFQ response format may enhance its utility.

Keywords: Online flow; Internet use; Psychometrics; Item Response Theory

Internet use has become an integral part of modern-day life, especially for adolescents across the globe [1,2,3,4]. This level of internet expansion has been partially attributed to technological advancements that have increased the vividness and interactivity of the medium [1,4,5]. In this context, the level of engagement experienced while an individual is using the internet appears to be pivotal, with different engagement levels being simultaneously associated with both the positive and negative effects of internet use [6-11].

Past literature posits that a user's level of online engagement is underpinned by the interplay of their personal characteristics, their surroundings, and features of the medium itself such as online flow [10,12-25]. Online flow has been defined as an individual's engagement/absorption in their online activity based on the progressive response to online demands, in a way that matches the rate of increase of an individual's performance to prevent boredom and maintain their consistent engagement [13-14]. Moreover, the escalation of an individual's online flow experience has been related to problematic behaviours such as Internet Gaming Disorder [8-12,15,22]. This has triggered the development of various instruments to assess the construct [13,18-19], including the five-item Online Flow Questionnaire (OFQ) [13].

Of the various online flow measures developed, the OFQ was chosen to be studied here for a number of compelling reasons, namely its: (i) theoretical correspondence with the flow construct considering elements/dimensions of challenge (OFQ Items 1, 4), merging action and awareness (OFQ Item 1), task concentration (OFQ Item 1), a sense of potential control (OFQ Item 6), loss of self-consciousness (OFQ Item 1), altered sense of time (OFQ Item 2) and autotelic experience (i.e., the joy lies within the activity it-self and not the activity's product (Items 1, 9); (ii) broadly accepted use; (iii) robust psychometric properties, (iv) brief structure (only five items, allowing it to easier incorporate in lengthier surveys); (v) distinct, yet complimentary content to other immersion/ transportation scales (i.e. it explicitly emphasizes one's engagement/immersion with what they do online, and not their virtual context or persona/avatar of representation within it in an undifferentiated manner) and; (vi) capacity to accommodate international comparisons due to its use in different cultural samples and adaptation across different languages (e.g., English, Greek, Cypriot; [10-13, 22-24]). Despite the scale's original validation using a predominantly adult sample of internet users (61% ranged between 20–40 years [12]), later studies confirmed its validity and reliability across different age-groups and national samples. However, metric invariance limitations have been highlighted, and the items' differentials have yet to be assessed [10,22-23]. Moreover, it is likely that higher online flow scores reported by males compared to females, may also be related to hypothesized different gender response patterns in addressing survey instruments assessing an individual's online behaviour [24-30]. Additionally, considering that online flow has been shown to play an important role in disordered online use behaviors (such as online gaming disorder), it may be valuable in identifying at-risk populations [22]. Therefore, there is a need to employ novel approaches such as Item Response Theory (IRT) to address these limitations.

Item Response Theory

IRT offers a unique framework to assess the psychometric properties of an instrument at the item and scale level [31]. Specifically, IRT models employ a non-linear logit function (Item Characteristic Curve; ICC) and parameter logistic (PL) such as discrimination (α) and difficulty (β) to explain how the likelihood of endorsing an item changes at different levels of the latent-trait (θ) [31-32]. Here, α evaluates the relationship between item and θ to describe its ability to differentiate θ levels [32]. Similarly, β refers to the minimum level of θ needed to endorse a particular item [32]. In this context, IRT models constraining α to be equal across items are considered 1PL, and models allowing free estimation of α across items are 2PL [31]. Additionally, IRT enables the estimation of θ-dependent reliability indices (conditional upon levels of θ), and investigation of differential item functioning (DIF) across groups of interest (e.g., males and females [33-34]). Finally, while IRT does not assume normality (unlike Classical Test Theory, CTT), it enables estimation of prevalence rates (i.e., ±2 SD) via the Summed Score Expected a Posteriori (SSEAP)[θ|x] [see 31, p.179;32]).

The present study

The present study contributes to the literature by considering the optimum measurement of online flow, alongside examining OFQ cut-off points [12-13]. This is done by first applying IRT analyses on the OFQ, while evaluating 1PL and 2PL models to examine the responses of a large sample of high-school students from Greece on the binary version of the OFQ scale. Specifically, it assesses discrimination, difficulty, and reliability of the OFQ (at both item and scale levels), as well as the DIF across males and females. Furthermore, identification of salient groups (i.e., beyond 2 SD) will be conducted via the SSEAP. The findings have significant implications regarding the assessment of online flow in research and client work (e.g., problematic internet use treatment field) by providing guidelines considering the structure of the construct and the psychometric properties of this instrument.

Methods

Participants

The data comprised the use of an archival dataset with a sample of 1579 students ($M_{age=}$16.12 years, SD=.85; 50.5% females) collected from Greek high schools. Although the participants' mean age difference across the two biological genders (females=16.1, SD=.81; males=16.2, SD=.98) was significant (*t*=-3.28, *p*=.001), the effect size was small (Cohen's *d*=-.177) [35]. Most participants were Greek (78.1%) and Albanian (12.8%), and 43.1% of those declared to be Massively Multiplayer Online gamers. Missing values considering the five OFQ items ranged between 1.1% for Item 1 and 3.7% for Item 2. These rates were below the recommended 5% missing value threshold [36].

Instruments

The Online Flow Questionnaire (OFQ) [12] assesses how much an individual is absorbed by the online activity using five questions with binary responses ("No"= 0 or "Yes"=1). Examples of items include *"In your Web navigation, have you ever experienced the feeling of 'time going too fast'?"*. Total possible scores range from 0-5, with higher scores representing higher levels of flow experience. The scale's internal reliability (classical test theory composite/scale score, not accounting for an individual's level of online flow experienced in IRT) was acceptable for the current data (ω=.74) [37]. The scale was administered in Greek after bi-directional translation by bilingual translators following standardized international guidelines [38]. More specifically, translators with two mother tongues, Greek and English, were divided into two groups. Group 1 first translated the OFQ from English to Greek and Group 2 back-translated the Group 1 version back to English. Minor differences detected between the Group 2 OFQ translation, and the original form were discussed and reconciled between the two groups to finalize the Greek version of the OFQ.

Procedure

After obtaining approval from the Institutional Human Research Ethics Committee, archival data collected between 2009-2012 in a paper–pencil format was accessed. Data were collected in class during two teaching hours after having received approvals by the: (i) Ministry of Education (ii) Principal and schoolteachers, and (iii) parents. The sample was situated within the extended Athens metropolitan area. Students were selected using the proportional stratified random sampling method, based on the Ministry of Education inventory card.

Statistical analyses

All IRT analyses were conducted with IRT PRO [39]. 1PL and 2 PL models were estimated, and $\Delta\chi^2$ based on $\chi^{2loglikelihood}$ was employed to identify the best fitting model [31-32]. Subsequently, model fit was determined by concurrently considering: (i) traditional fit indices ($\chi^{2Loglikelihood}$, and G^2 [40]); (ii) marginal likelihood information statistics (using one and two-way marginal tables to correct for potential sparsity) M_2; and (ii) the RMSEA (.08 and lower=sufficient fit) [41-42]. However, given potential sensitivity of M_2 and χ^2 to large sample sizes (N>1000), emphasis was given to RMSEA to assess goodness of fit [42]. Considering α, the following rates/ranges were taken into consideration: 0=non discriminative; 0.01-0.34=very low; 0.35-0.64=low; 0.65-1.34=moderate; 1.35-1.69=high; and >1.70=very high [51]. At the scale level, test reliability was assessed with the test information function (TIF) and the overall test performance via the test characteristic curve (TCC) [42]. Additionally, DIF statistics using Wald tests were employed to identify potential psychometric differences across traditional gender groups (with *p*<0.5 as the significance level for non-invariance). Subsequently, to correct for potential type 1 error inflation, invariant items were anchored testing only non-invariant items [33]. Finally, the conversion of the OFQ raw scores into online flow experience levels was conducted based on the optimum fit IRT model conversion table [31].

Results

First, IRT assumptions of unidimensionality, local dependency, and monotonicity were tested [31]. A confirmatory factor analysis (CFA) was conducted with R Studio (Lavaan package; [43]) with FIML (full information maximum likelihood estimator) due to having missing values [32] to test OFQ's factorial structure's properties. Unidimensionality was assumed as the OFQ showed acceptable fit indices (χ^2=20.7, p<.001, CFI=.949, RMSEA=.063; see Supplementary Material). Local independence was assumed given that all pairs of items showed LD $\chi^{2<}$10 [44]. Finally, monotonicity was examined through inspection of higher levels of difficulty for "yes" responses for each of the items (see Table 2), as well as the visual inspection of the TCC for the whole scale (see *Figure 4*).

Therefore, IRT model estimation was progressed using the Bock-Aitkin marginal maximal likelihood algorithm with expectation-maximization [45]. While the 1PL model showed acceptable fit ($\chi^{2\text{Loglikelihood}=}$8418.35; G^2 [25]=55.75; χ^2 [25]=60.83; $M_{2=}$23.94; RMSEA=.03), the 2PL model showed improved fit ($\chi^{2\text{Loglikelihood}=}$8412.75; G^2 [25]=50.48; χ^2 [21]=52.25, p<.001, $M_{2=}$20.15; RMSEA=.03), although marginally ($\Delta\chi^{2=}$8.58, p=.073; Table 1). Given the considerable variation in α when free estimation across items was allowed, the 2PL model as optimum fit was identified.

-Table 1-

Considering α, items ranged between 1.10 (Item 2) and 1.51 (Item 4). Similarly, factor loadings ranged between .54 (λ Item 2) and .66 (λ Items 3 and 4) [45]. Items' discrimination power descended in the following order: 4, 3, 1, 5, and 2 (see Table 2). Considering β, there was a considerable level of variation, resulting in the following descending item sequence 'Item 5', 'Item 1', "Item 4', "Item 3', and "Item 2' (see Table 2/*Figure 1*).

-Table 2/*Figure 1*-

Considering the items' reliability across the different levels of the latent-trait, meaningful variations were similarly confirmed. More specifically, item 1 provided the highest level of information/reliability in the range between -0.5 SD and +2.5 SD beyond the mean (see Item Information Function, IIF; *Figure 2*). Item 2 provided considerably higher information in the area between -3 SD and +1 SD. Item 3 gave more reliable information for respondents in the area between -2.5 SD and +2 SD. Item 4 information quality/reliability were higher in the area of -1.5 SD and +2 SD. Item 5 provided better information in the area between +1 SD and +2 SD.

-Figure 2-

Considering the information provided by the scale as a whole, improved information (Test Information Curve [TIC]) scores were around -1 SD and + 2 SD (*Figure 3*). Furthermore, the TCC illustrates that the level of online flow experienced, as per total score reported, increased (in particular) from -1 SD for a raw score of 1, to +2 SD for a raw score of 5 (*Figure 4*). These results suggest that the whole scale provides a reliable psychometric measure for assessing online flow experiences. Nevertheless, as indicated by the different item information curves, as well as the TIC, although the scale sufficiently differentiates those who had or had not had flow experiences, it fails to sensitively differentiate between those with lower levels, as well as those with higher levels (e.g., "low" to "very low"; 'high' to very 'high').

-Figure 3/Figure 4-

Considering differential functioning of the OFQ items between males and females, items 1, 3, and 4 were non-invariant (see Table 3) and presented significantly different between-groups (total χ^2 p<.05). Therefore, the two invariant items were then anchored, and the differential functioning of the non-invariant items was re-calculated to address familywise type I error [33]. When items 2 and 5 were anchored, only item 3 presented a significant difference between groups (total χ^2=15.90, df=2, p<.001, difficulty $\chi^2{}_{\text{cja}}$=15.60, df=1, p<.001). This is illustrated in *Figure 5*, where males (see Group 2) present with higher probability of endorsing a positive response, indicating that males are more likely to endorse item 3 compared to females.

-*Figure 5*, Table 3-

Lastly, the raw OFQ score, at a level of 2SDs above the mean of the latent-trait level, was translated to equal or above a score of 5. Based on this, it could be suggested as a conditional (before clinical assessment confirmation) diagnostic cut-off point for those who have experienced flow. Table 4 presents the OFQ raw scores and their translation into online flow latent scores (EAP[θ|x]), as well as the proportion of the sample reflected by each rate (Modeled Proportion).

-Table 4-

Discussion

The present study is the first to assess the psychometric properties of the OFQ at both item and scale levels using a large and normative sample from Greece exceeding 1000 participants. This was done by employing IRT procedures to assess items' α and β [31-32]. Findings demonstrated that the OFQ items differed across all the psychometric properties assessed, enabling consideration of a potential ranking of item responses in assessments. The analyses also indicated that the reliability provided by the scale as a whole was rather low for both the low and the high extremes of online flow experiences reported (± 2SD beyond the mean). Overall, the findings suggest that although the items' content may be appropriate for assessing the latent construct of online flow, the addition of more items and/or the use of Likert scale responses with more than two options may be considered [47]. The latter is particularly important if researchers wish to accurately capture variations of lower and higher flow experiences, which could be significant in the context of problematic internet use over time.

Scale and items IRT properties

Considering α, non-significant variation was observed between the 1PL model and the 2PL model described here. Nevertheless, all OFQ items ranged between moderate and high levels of α. When considering the mild α differences regarding OFQ shown in the 2PL model, items related to intrinsic motivation experienced in relation to the activity (items 3 and 4) may be stronger in this respect. Alternatively, feelings related to loss of control and a distorted sense of time appear to have relatively lower α (items 2 and 5). This aligns with similar findings in relation to relevant behaviors (see IGD) [30]. Alternatively, given that items reflecting an individual's flow pleasure-seeking motivation demonstrated higher α, they may need to be considered as a priority when assessing online flow (see OFQ items 4 and 3).

Item variations were also confirmed in relation to β, with a descending item sequence (items 5, 1, 4, 3, and 2). Specifically, item 5's low α may be explained by its high difficulty.

This suggests that because a positive response requires a higher level of the latent-trait (higher β), variations below that level may not be adequately identified (lower α). Interestingly, item 1 involving the description of the online flow experience, appears to combine a moderate α with a relatively higher β compared with other items. This is somewhat expected, as it presents more inclusive/descriptive of the various flow aspects, therefore balancing between β and α [12,14]. Last, item 2 was the worst performing OFQ item based on both α and β. This aligns with findings in relation to the similar low performance of "escapism" items in IGD [30], suggesting that a 'distorted sense of time while being online' constitutes a rather common experience among internet users.

Variations were also noticed regarding the items' reliability indices. Item 1 Information Function (IIF) exhibited the highest level of information/reliability for participants in the range between -0.5 SD and +2.5 SD beyond the mean, being the most reliable item. Additionally, item 2 provided considerably higher information between -3 SD and +1 SD, being a more appropriate question to ask for those with lower levels of online flow experience. Finally, items 3 and 4 showed more reliability among participants in the 'normative' area (-2.5 SD to +2 SD) suggesting that these items may need to be considered more cautiously for those with significantly lower and/or higher levels of flow experience.

The appropriate reliability indices provided by OFQ items in the average range (with the exception of item 2), explain the reliable information provided by the scale as a whole, with improved information scores situated around -1 SD and +2 SD. This suggests that the OFQ is more reliable for participants within the average range of online experiences, but less reliable for those situated in both the high and low extremes. As suggested by Embretson and Reise [31], increasing the number of items may not necessarily result in increased reliability indices because CTT and IRT perform substantially different in this respect. Therefore, 'adaptive' forms of scales should be employed prioritizing items that correspond with participants' latent-trait levels [31].

Last, regarding DIF, findings suggested that item 3 response patterns were significantly different between males and females. This indicates that males (compared to females) require a lower level of OFQ to 'enjoy' spending time online. In line with previous studies, it is possible that males' higher achievement/challenge orientation and computer use inclinations, cultivated through socialization in more conservative societies (such as the Greek society may be compared to other Western countries) contribute to this difference [21-22,27]. Nevertheless, further cross-cultural investigation is warranted to clarify this interpretation.

OFQ raw scores and flow experience

Considering the translation of the raw OFQ scores into latent online flow levels two main findings can be highlighted. First, the level of online flow experienced, as per total score reported, increased (in particular) from -1 SD below the mean for a raw score of 1. Second, for an individual to be considered experiencing an excessive online flow (i.e., over +2 SD), a raw OFQ score of 5 is required. These findings reinforce the conclusion that although the content of the OFQ items chosen is relevant, item addition and/or changes in the response options may be required [47]. Moreover, scale reliability related limitations (see TIF) dictate caution should be applied regarding the conversion of raw into scaled IRT scores, for very low and very high OFQ levels.

Despite the relevance of the current OFQ items supported by the findings, the limited number of questions, which may not adequately capture the subtleties of online flow, should be acknowledged. Thus, further (theory-based) questions/items could be added to enrich the present scale, provided they are also assessed regarding their psychometric properties. In that context, one should consider applying the IRT analysis approach adopted in the present study, as it offers valuable psychometric information. This might be utilized either in improving the present instrument and/or informing other measures assessing one's engagement/absorbance by their internet activity.

Limitations, further research directions, and conclusions

The present study combined two significant strengths (i.e., utilizing a large sample and employing IRT analyses). This allowed exploration of the psychometric properties of the OFQ at both item and scale level considering the discrimination, difficulty, and reliability capacities. Despite such strengths, limitations need to also be acknowledged. First, the archival data analyzed refers to a specific cultural population and therefore restricts the generalizability of the findings. Second, the lack of multidimensionality related to the OFQ may hinder insight regarding the accuracy of assessing varying online flow aspects suggested [18,20]. Third, given the non-significant drop of fit between the 1PL and 2PL model, the 1PL model may also be applicable. Fourth, OFQ scores may need to be supplemented by use of interview-related assessment, or further explored with novel methods such as network analysis [48] to better understand an individual's accurate level of online flow experienced. Fifth, the current OFQ composition and items content/phrasing may need to be further improved and diversified to more accurately assess online flow across different online applications (i.e., games, social media) and research contexts (i.e., blended/augmented vs. exclusively online) in the contemporary broad and varied online environment. In that line, it is likely that the limited number of the current OFQ questions/items, may not adequately capture the subtleties of online flow (and thus more/different items/questions should be added/tested). Future research should take into account these limitations to advance the area.

Moreover, further elaboration of terms employed within OFQ items (e.g., Item 4, "positive challenge") may be achieved by either incorporating specific examples and/or allowing to participants to provide more details, possibly by utilizing open descriptive questions. For instance, a qualitative accompanying item requesting an individual to describe the experience related to their "yes" response (regarding each of the five primary binary items), may allow better capturing the multiple flow dimensions [13]. Additionally, measurement invariance and longitudinal measurement invariance studies need to be conducted to investigate likely psychometric variations of the scale's properties and its items across genders, age ranges, cultural groups, and over time. Finally, comparative psychometric studies of the OFQ (as a scale exclusively focusing on an individual's engagement with their online activity) with other available scales developed that assess similar constructs (e.g., virtual context/avatar immersion, transportation [49]), might be useful in solidifying concurrent and convergent validity. Such future findings would increase the robustness of the measurement and the empirical evidence reflecting the significant effects of online flow.

Despite such limitations, the present study outlines three significant findings: (i) OFQ items perform differently with respect to discrimination and difficulty although all items are generally within acceptable ranges; (ii) the OFQ as whole performs better and is more reliable for individuals between the moderately average and the high average range; and (iii) the inclusion of a Likert scale response format with more options would need to be considered for the scale to better capture the variability in the higher ranges of the online flow experience.

Declarations

Acknowledgments: The authors would like to acknowledge the team effort making possible the production of this manuscript.

Authors confirmation/contribution: VS contributed to the article's conceptualization, data curation, formal analysis, methodology, project administration, and writing of the original draft. TF contributed to data curation, writing of the original draft, review, editing the final draft and project administration. DZ contributed to the article's conceptualization, data curation, formal analysis, methodology, project administration, and writing of the original draft. MP and MG contributed to writing, reviewing, and editing the final draft.

Competing Interests: All authors have no competing interests.

Funding: The data collection and processing for the present study was obtained through a grant from The Victoria University, Early Career Researcher Fund ECR 2020, number 68761601. The Australian Research Council, Discovery Early Career Researcher Award, 2021, number DE210101107.

References

1. Anderson EL, Steen E, Stavropoulos V. Internet use and problematic internet use: A systematic review of longitudinal research trends in adolescence and emergent adulthood. *International Journal of Adolescence and Youth* 2017;22(4):430-54.
2. Kao G, Joyner K., Balistreri, KS. *The company we keep: Interracial friendships and romantic relationships from adolescence to adulthood*. Russell Sage Foundation; 2019.
3. Internet World Stats. *World Internet Users Statistics and 2020 World Population Stats. Internetworldstats.com*. Retrieved 8 August 2020, from https://www.Internetworldstats.com/stats.htm.
4. Ragnedda M. *The third digital divide: A Weberian approach to digital inequalities.* Routledge; 2017.
5. Scheerder A, Van Deursen A, Van Dijk J. Determinants of Internet skills, uses and outcomes. A systematic review of the second-and third-level digital divide. *Telematics and informatics* 2017;34(8):1607-24.
6. Salmela-Aro K, Muotka J, Alho K, Hakkarainen K, Lonka K. School burnout and engagement profiles among digital natives in Finland: A person-oriented approach. *European Journal of Developmental Psychology* 2016;13(6):704-18.
7. Alexandraki, K., Stavropoulos, V., Burleigh, T. L., King, D. L., & Griffiths, M. D. (2018). Internet pornography viewing preference as a risk factor for adolescent internet addiction: The moderating role of classroom personality factors. *Journal of Behavioral Addictions, 7*(2), 423-32.
8. Cash H, Rae CD, Steel AH, Winkler, A. Internet addiction: A brief summary of research and practice. *Current Psychiatry Reviews* 2012;8(4):292-8.
9. Stavropoulos V, Burleigh TL, Beard CL, Gomez R, Griffiths MD. Being there: a preliminary study examining the role of presence in internet gaming disorder. *International Journal of Mental Health and Addiction* 2019;17(4):880-90.
10. Hu E, Stavropoulos V, Anderson A, Scerri M, Collard J. Internet gaming disorder: Feeling the flow of social games. *Addictive Behaviors Reports* 2019;9:100140.
11. American Psychiatric Association. *Diagnostic and statistical manual of mental disorders* (5th ed.). APA; 2013Stavropoulos V, Burleigh TL, Beard CL, Gomez R, Griffiths MD. Being there: a preliminary study examining the role of presence in internet gaming disorder. *International Journal of Mental Health and Addiction* 2019;17(4):880-90.

12. Chen H, Wigand RT, Nilan MS. Optimal experience of web activities. *Computers in Human Behavior* 1999;15(5):585-608.
13. Nakamura J, Csikszentmihalyi M. *The concept of flow. InFlow and the foundations of positive psychology.* Springer; 2014.
14. Stavropoulos V, Alexandraki K, Motti-Stefanidi F. Flow and telepresence contributing to internet abuse: Differences according to gender and age. *Computers in Human Behavior* 2013;29(5):1941-8.
15. Kelley TM, Pransky J, Lambert EG. Realizing improved mindfulness/flow/mental health through understanding three spiritual principles. *Journal of Spirituality in Mental Health* 2017;19(2):133-50.
16. Sanjamsai S, Phukao D. Flow experience in computer game playing among Thai university students. *Kasetsart Journal of Social Sciences* 2018;39(2):175-82.
17. Liu CC. A model for exploring players flow experience in online games. *Information Technology & People* 2017.
18. Webster J, Trevino LK, Ryan L. The dimensionality and correlates of flow in human-computer interactions. *Computers in Human Behavior* 1993;9(4):411-26.
19. Hoffman DL, Novak TP. Flow online: lessons learned and future prospects. Journal of Interactive Marketing 2009;23(1):23-34.
20. Sicilia M, Ruiz S. The role of flow in web site effectiveness. *Journal of Interactive Advertising* 2007;8(1):33-44.
21. Stavropoulos V, Griffiths MD, Burleigh TL, Kuss DJ, Doh YY, Gomez R. Flow on the Internet: a longitudinal study of Internet addiction symptoms during adolescence. *Behaviour & Information Technology* 2018;37(2):159-72
22. Hu E, Stavropoulos V, Anderson A, Clarke M, Beard C, Papapetrou S, Gomez R. Assessing online flow across cultures: A two-fold measurement invariance study. *Frontiers in Psychology* 2019;10:407.
23. Jamshidi D, Keshavarz Y, Kazemi F, Mohammadian M. Mobile banking behavior and flow experience: An integration of utilitarian features, hedonic features and trust. *International Journal of Social Economics* 2018; 45(1):57-81.
24. Weber R, Tamborini R, Westcott-Baker A, Kantor B. Theorizing flow and media enjoyment as cognitive synchronization of attentional and reward networks. *Communication Theory. 2009* Nov 1;19(4):397-422.
25. Wu TY, Lin CY, Årestedt K, Griffiths MD, Broström A, Pakpour AH. Psychometric validation of the Persian nine-item Internet Gaming Disorder Scale–Short Form: Does gender and hours spent online gaming affect the interpretations of item descriptions? *Journal of Behavioral Addictions* 2017;6(2):256-63.
26. Stavropoulos V, Mastrotheodoros S, Burleigh TL, Papadopoulos N, Gomez R. Avoidant romantic attachment in adolescence: Gender, excessive internet use and romantic relationship engagement effects. *PloS One* 2018;13(7):e0201176.
27. Stavropoulos V, Gomez R, Motti-Stefanidi F. Internet gaming disorder: A pathway towards assessment consensus. *Frontiers in Psychology* 2019;10:1822.
28. Stavropoulos V, O'Farrell DL, Baynes KL, Pontes HM, D Griffiths M. Depression and disordered gaming: does culture matter? *International Journal of Mental Health and Addiction* 2020.
29. Stavropoulos V, Mastrotheodoros S, Papapetrou S, Gomez R, Beard C, Motti-Stefanidi F. Measurement invariance: The case of measuring romantic attachment in Greek and Cypriot adolescents. *European Journal of Developmental Psychology* 2019;16(3):362-71.

30. Gomez R, Stavropoulos V, Beard C, Pontes HM. Item response theory analysis of the recoded internet gaming disorder scale-short-form (IGDS9-SF). *International Journal of Mental Health and Addiction* 2019;17(4):859-79.
31. Embretson SE, Reise SP. *Item response theory*. Psychology Press; 2013.
32. De Ayala RJ. *The theory and practice of item response theory*. Guilford Publications; 2013.
33. Zarate D, Marmara J, Potoczny C, Hosking W, Stavropoulos V. Body Appreciation Scale (BAS-2): measurement invariance across genders and item response theory examination. *BMC Psychology* 2021;9(1):1–15.
34. Marmara J, Zarate D. Functionality Appreciation Scale (FAS): Item Response Theory Examination. *BMC Psychology*, preprint 10.21203/rs.3.rs-1148688/v1
35. Mahadevan L. *The effect size statistic: Overview of various choices*. 2000.
36. Little RJ. A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association* 1988;83(404):1198-202.
37. Viladrich C, Angulo-Brunet A, Doval E. A journey around alpha and omega to estimate internal consistency reliability. *Annals of Psychology* 2017;33(3):755-82.
38. Beaton D, Bombardier C, Guillemin F, Ferraz MB. *Recommendations for the cross-cultural adaptation of health status measures.* New York: American Academy of Orthopaedic Surgeons. 2002 Mar;12:1-9.
39. Cai L, Du Toit SH, Thissen D. *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]*. Chicago, IL: Scientific Software International. 2011.
40. Orlando M, Thissen, D. Likelihood-based item fit-indices for dichotomous item response theory models. *Applied Psychological Measurement* 2000;24(1):50-64.
41. Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling* 1999;6(1):1-55.
42. Stone CA, Zhang B. Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement* 2003;40(4):331-52.
43. Rosseel Y. Lavaan: An R package for structural equation modelling. *Journal of Statistical Software* 2021;48(2):1-36.
44. Chen WH, Thissen D. Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics* 1997;22(3):265-289.
45. Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* 1981;46(4):443-459.
46. Thompson B. Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools* 2007;44(5):423-32.
47. Linacre JM. Optimizing rating scale category effectiveness. *Journal of Applied Measurement* 2002;3(1):85-10
48. Zarate D, Ball M, Montag C, Prokofieva M, & Stavropoulos V. Unravelling the web of addictions: A network analysis approach. *Addictive Behaviors Reports* 2022; 15:100406.
49. Appel M, Gnambs T, Richter T, Green MC. The Transportation Scale–Short Form (TS–SF). *Media Psychology* 2015;18(2):243-66.

Table 1. *1PL IRT and 2PL IRT model fit comparison*.

Statistics based on the loglikelihood		
Model	2PL	1 PL
-2loglikelihood:	8412.75	8418.4
Statistics based on the full item classification		
Model	2PL	1PL
G^2	50.48	55.75
	DF=21, *p*=.0003	DF=25, *p*=.0004
	RMSEA= .03	RMSEA=.03
χ^2	52.25	60.83
	DF=21, *p*=.0002	DF=25, *p*=.0001
	RMSEA= .03	RMSEA= .03
$\Delta\chi^2$	8.58	
	ΔDF=4, *p*=.073	
Statistics based on one- and two-way marginal tables		
Model	2PL	1PL
M_2	20.15	23.94
	DF=5, *p*=.0012	DF=9, *p*=.0044
	RMSEA= .04	RMSEA= .03

Table 2. *OFQ Item 2PL IRT Properties*

Item	*α*	*β*	*λ (Loadings)*	*S*- χ^2	df	*p*
1	1.37	1.03	0.63	5.19	3	0.1579
2	1.10	-1.47	0.54	7.91	3	0.0479
3	1.50	-0.40	0.66	5.76	3	0.1235
4	1.51	0.15	0.66	7.35	3	0.0615
5	1.17	1.74	0.57	8.85	3	0.0313

Note. *α* defines the capacity of an item to discriminate between varying levels of online flow (θ). The β defines the level of online flow intensity, where subsequent response rates are more probable to be positive. The λ defines the amount of variance of an item explained by the latent factor. The S- χ^2 describes the item-fit statistic for each item and behaves similarly to χ^2 in CFA, with insignificant rates showing no deviation of the item modelling from the data.

Table 3. Differential Item Functioning (DIF) Statistics for Graded Items

Group 1	Group 2	Total χ^2	*d.f.*	*p*	χ^2_{a}	*d.f.*	*p*	$\chi^2_{c\mid a}$	*d.f.*	*p*
1	1	9.6	2	0.0083	1.1	1	0.3041	8.5	1	0.0035
2	2	0.0	2	0.9921	0.0	1	0.9026	0.0	1	0.9773
3	3	23.0	2	0.0001	0.3	1	0.5848	22.7	1	0.0001
4	4	10.2	2	0.0059	0.3	1	0.6084	10.0	1	0.0016
5	5	0.3	2	0.8696	0.3	1	0.6041	0.0	1	0.9194

Table 4. *Summed Score to Scale Score Conversion*

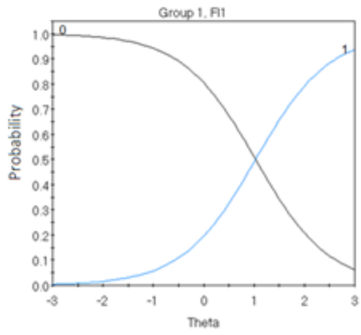
Summed Score	EAP[θ|x]	SD[θ|x]	Modelled Proportion
0	-1.301	0.713	0.0923016
1	-0.728	0.667	0.2170774
2	-0.144	0.635	0.2602315
3	0.416	0.631	0.2346682
4	0.967	0.653	0.1451999
5	1.533	0.692	0.0505214

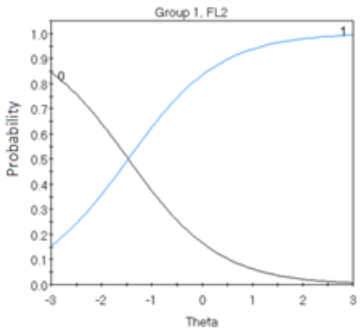
Fig. 1 Items’ Characteristic Curves (ICC). These plots demonstrate how the probability of endorsing an OFQ item (x axis) change as levels of the latent trait change (y axis). A 0 denotes a negative response, and a 1 denotes a positive response to the item.

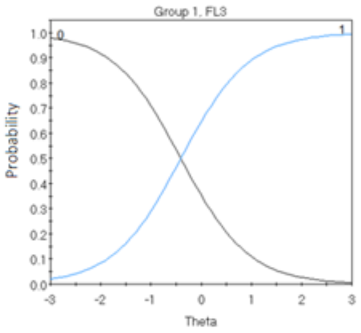
Fig. 2 Item Information Function (IIF). These plots demonstrate how reliability indices vary (x axis) with changes in the latent trait (y axis), with higher reliability index representing more information.

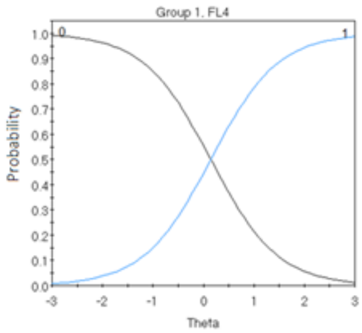
Fig. 3 Test Information Function (TIF; left panel) and Test Characteristic Curve (TCC; right panel). The TIF demonstrates the relationship between standard errors and reliability indices (i.e., smaller standard errors provide more information). The TCC shows expected OFQ scores as a function of latent trait levels.

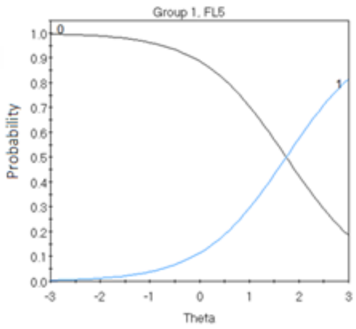
Fig. 4 Differential Item Functioning (DIF) for non-invariant items across males (Group 1) and females (Group 2). As observed here, men require lower levels of the latent trait (*θ*) to endorse Item 3.

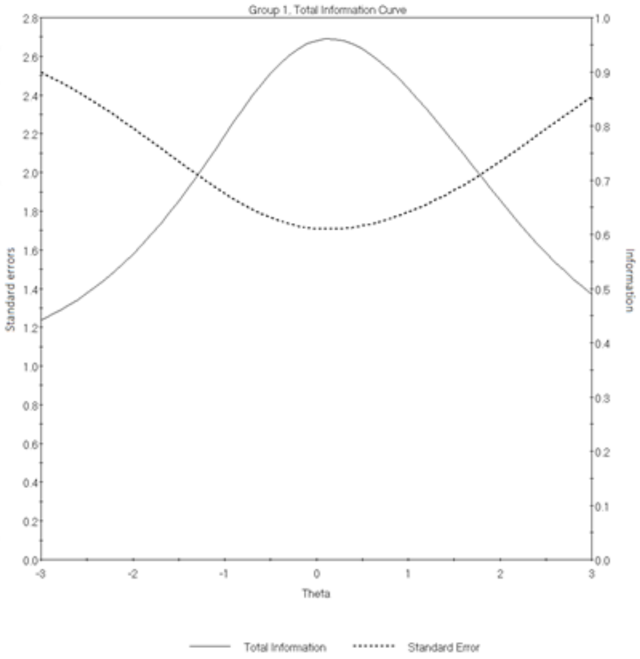

Group 1, FI1
Probability
Theta
0
1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3

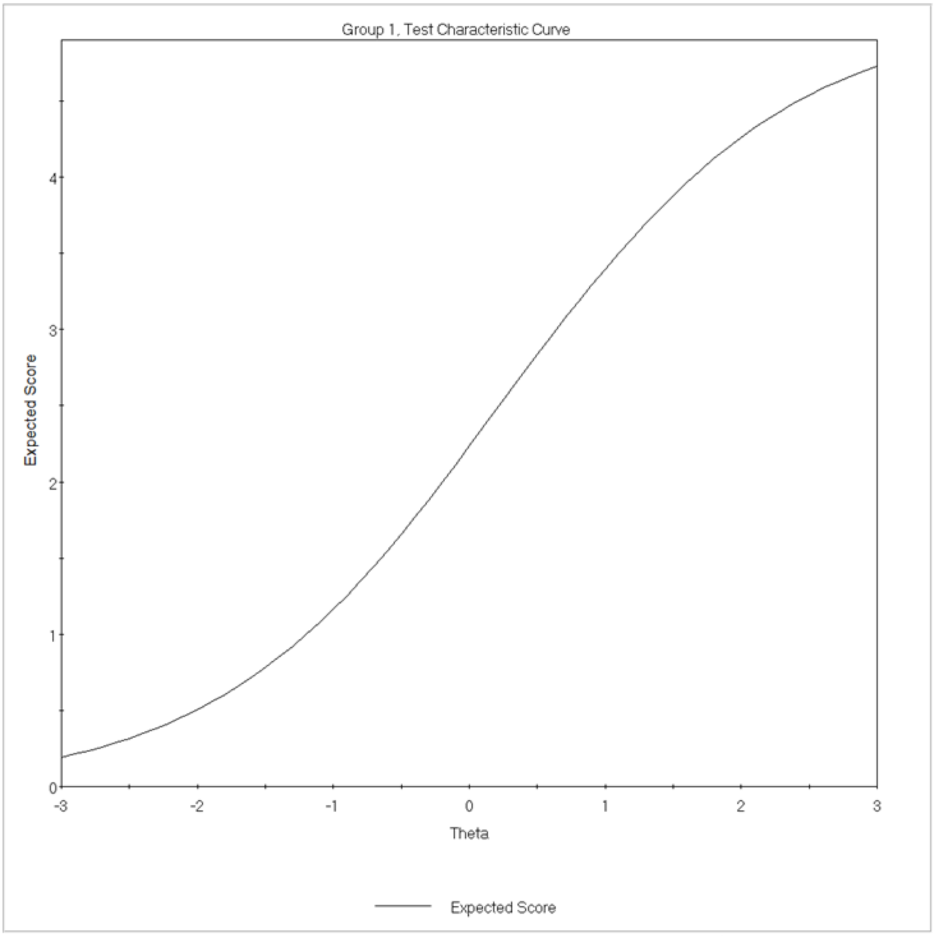

Group 1, FL2
Probability
Theta
0
1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3

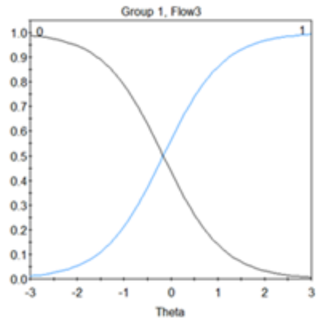

Group 1, FL3
Probability
Theta
0
1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3

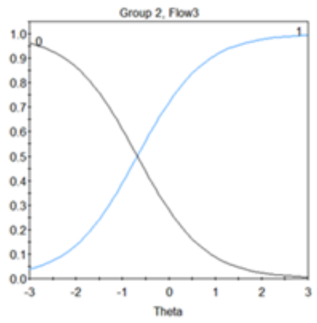

Group 1, FL4
Probability
Theta
0
1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3

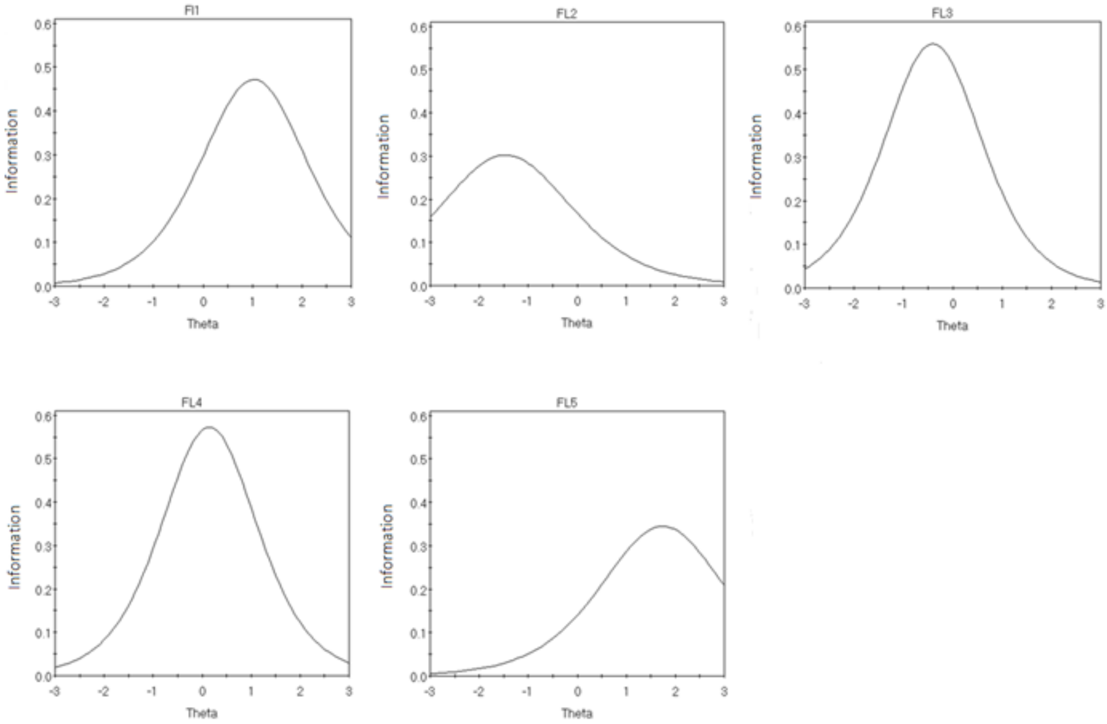

Group 1, FL5
Probability
Theta
0
1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3


Group 1, Total Information Curve
Standard errors
Information
Theta
2.8
2.6
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3
Total Information
Standard Error


Group 1, Test Characteristic Curve
Expected Score
0
1
2
3
4
-3
-2
-1
0
1
2
3
Theta
Expected Score


Group 1, Flow3
0
1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3
Theta


Group 2, Flow3
0
1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3
Theta



Fl1
FL2
FL3
FL4
FL5
Information
Theta
0.0
0.1
0.2
0.3
0.4
0.5
0.6
-3
-2
-1
0
1
2
3