**Identifying Gaming Disorders by Ontology:**

**A Nationally Representative Registered Report**

[Stage 1, v5]

Veli-Matti Karhulahti1\*,

Jukka Vahlo1, 4,

Marcel Martončik2,

Matti Munukka1,

Raine Koskimaa1,

Mikaela von Bonsdorff1, 3

1 University of Jyväskylä, Jyväskylä, Finland

2 University of Prešov, Prešov, Slovakia

2 Institute of Social Sciences, CSPS SAS, Slovakia

3 Folkhälsan Research Center, Helsinki, Finland

4 University of Turku, Turku, Finland

\* Corresponding author, vmmkar@utu.fi

**Author Note**

Review version 2.0, 12.1.2022. This paper has not been peer reviewed.

**Word count:** 4331 (excluding abstract, captions, and references)  
**Appendix 1:** Pilot

**Appendix 2:** Research questions and hypotheses

**Materials**: <https://osf.io/6fqm5/>

**Funding**: The work was supported by the Finnish Work Environment Fund (200349), Academy of Finland (312397), and the Slovak Research and Development Agency under the Contract no. APVV-18-0140.

**Abstract**

Gaming-related health problems have been researched since the 1980s with numerous different “ontologies” as reference systems, from self-assessed “game addiction” to “pathological gambling” (in the DSM-IV), “internet gaming disorder” (in the 3rd section of the DSM-5) and most recently “gaming disorder” (in the ICD-11). Their differences have not been studied explicitly, however. In this registered report, we ask: how do screening instruments that derive from different ontologies differ in identifying associated problem groups? By using four central screening instruments, each of which represent a different ontological basis, we provide a nationally representative (*N*=8000) comparative “gaming disorder” prevalence (in Finland) and test three hypotheses concerning the possible multiplicity of the construct, which is currently discussed as a single “gaming disorder” in medical and scholarly domains.

**Introduction**

A lot of scientific effort has recently been invested in studying the relationship between technology use and wellbeing. A key theme in this research are the suggested mental disorders related to technology use, which outside diagnostic manuals are often discussed as “addictions,” such as “social media addiction.” At the time of writing, one technology-related disorder has received approval from the World Health Organization (WHO): in 2019, “gaming disorder” was confirmed to be included in the *International Classification of Diseases* (ICD-11)as a disorder due to addictive behaviors. Unlike the WHO, the American Psychiatric Association (APA) decided not to include such disorders for diagnosis in their *Diagnostic and Statistical Manual of Mental Disorders* (DSM-5), because the literature “suffers from a lack of a standard definition from which to derive prevalence data [and] understanding of the natural histories of cases” (APA 2013, p. 796). “Internet gaming disorder” was listed in the third section of DSM-5 as needing more research.

Despite “gaming disorder” now being part of the ICD-11, the problems addressed by the APA—how to identify and epidemiologically measure those with clinically significant problems—remain debated. Henceforth, we use *gaming-related health problems* (GRHP) as a reference to all constructs ("videogame addiction," “gaming disorder,” etc.) that relate to adverse mental, physical, and social aspects of gaming. A recent review of English GRHP screening instruments found 32 unique scales (King et al. 2020), and the scholars concluded five of them—including GAS7 (Game Addiction Scale: Lemmens et al. 2009) and IGDT10 (Internet Gaming Disorder Test: Király et al. 2017)—to have high evidential support. A later review found the prevalence rates of these five instruments to range from less than 1% to more than 14% in different age groups (Stevens et al. 2020). Despite scholars’ continuous attempts to make sophisticated epidemiological estimations, the fundamental problem remains unsolved: with at least 32 English screening tools that derive from diverse *ontological* grounds, what varies may not only be the prevalence rates, but also the constructs that are *being measured*. For instance, GAS7 is based on the definition of *pathological* *gambling* in the DSM-IV (seven symptoms four of which must be present), whereas IGDT10 is based on “internet gaming disorder” in the DSM-5 (nine symptoms five of which must be present). After “gaming disorder” was confirmed to be included in the ICD-11, new instruments now rely on the WHO’s ontology: three novel criteria that must all be met for establishing a diagnosis. Taken together, these differences reflect two dimensions of ontological diversity:

1. What criteria define the disorder? E.g., the DSM-5 and the ICD-11 list different criteria.
2. How criteria define the disorder? E.g., the DSM-5 disorder is diagnosable when some criteria are met, but the ICD-11 demands meeting all criteria.

Previous research has found many participants who express GRHP symptoms based on the DSM-5 not to have general health problems (Colder Carras & Kardefelt-Winther 2018), whereas some scholars further distinguish between the *degrees* to which criteria, symptoms, and health problems manifest (Myrseth & Notelaers 2018). Such investigations imply that it could be useful to distinguish more than one (not necessarily diagnostic) GRHP construct. Accordingly, our research question is:

**RQ**: How do screening instruments that derive from different ontological understandings differ in identifying GRHP groups?

The research question is investigated from three perspectives—prevalence, overlap, and health—which form three sub research questions.

**RQ-A**: How do GRHP screening instruments that derive from separate ontological understandings differ in *their prevalence rates (how many)*?

**RQ-B**: How do GRHP screening instruments that derive from separate ontological understandings differ in *who they identify (what characteristics)*?

**RQ-C**: How do GRHP screening instruments that derive from separate ontological understandings differ inthe *health of their identified groups (how healthy)*?

To formulate informed hypotheses, we carried out a pilot (*N*=1000) by using four central GRHP screening instruments with separate ontological foundations.

GAS7 (Lemmens et al. 2009): based on pathological gambling, as defined in the DSM-IV.

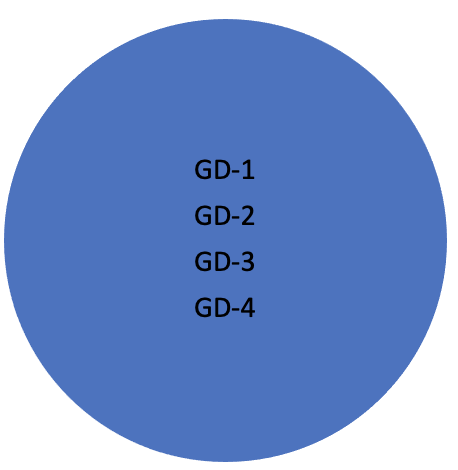
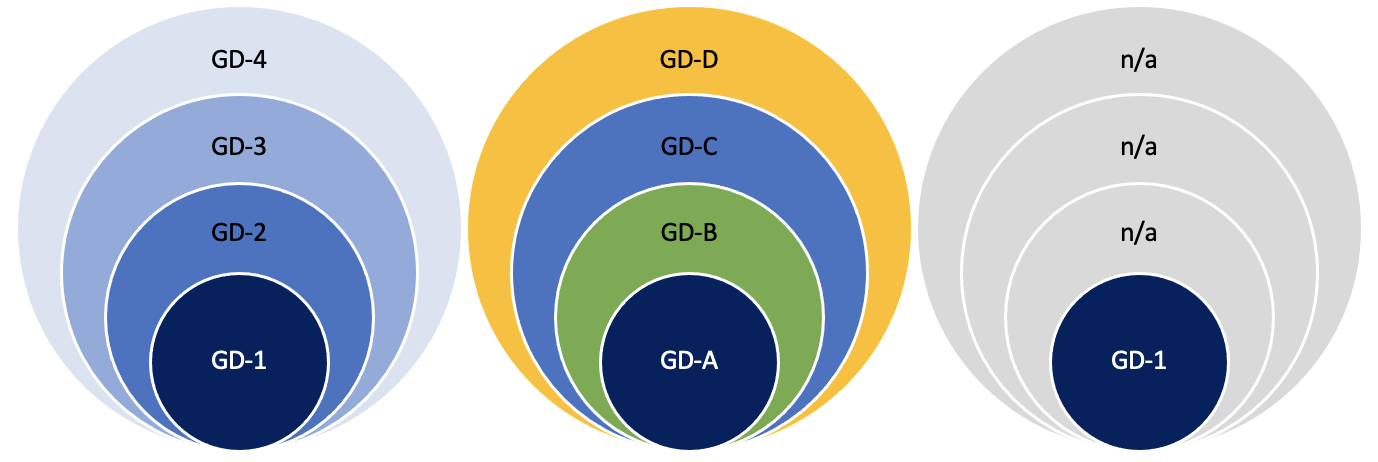
IGDT10 (Király et al. 2017): based on “internet gaming disorder,” as defined in the DSM-5.

GDT (Pontes et al. 2019): based on “gaming disorder,” as defined in the ICD-11.

THL1 (Salonen & Raisamo 2015): based on Self-assessed gaming problems.

Whereas all the above instruments measure GRHPs, the discrepancies in the above-cited literature imply that the construct(s) being measured might not be a single entity, but plural. Although our methods cannot investigate the clinical or etiological *nature* of these potentially multiple constructs, we can seek evidence for the *multiplicity* of the measured construct(s). Thus, we use “constructs” as references to those latent entities, which are shared by the (expectedly) different four groups identified by each of the four instruments, respectively. Accordingly, our study will pursue evidence for an improved conceptual organization of what the ICD-11 and many scholars today refer to as a single construct, “gaming disorder.”

The current epidemiological and historical understandings of “gaming disorder” are based on a long lineage of research throughout which multiple GRHPs have been treated as one and the same construct. In practice, studies on the literal “gaming disorder” (by the ICD-11) draw from the literature on “internet gaming disorder” (in the DSM-5), which in turn was built on studies on “gaming addiction” (pathological gambling in the DSM-IV) and others, such as “problem gaming,” that vary in their ontology by each study and research team. Sound evidence for this issue are the epidemiological reviews, which systematically mix gambling-based (DSM-IV), internet-based (DSM-5), general (ICD-11), and other ontologically diverse findings of GRHPs (e.g., Stevens et al. 2020). If we find evidence for our four ontologically diverse instruments to be similar in how they operate, this supports the *single-construct* idea of “gaming disorder” (which may manifest in many ways, Figure 1). If the instruments differ in how they operate, this indicates that efforts should be directed toward assessing the clinical (ir)relevance of *multiple* *constructs* (Figure 2).



**Figure 1**. Four exemplary scenarios where GRHP screening instruments identify overlapping groups. **On the left**: an example scenario where four screening instruments identify four different sized groups and they all capture the same construct on different severity levels (e.g. GD-1: severe problems, GD-4: mild problems). **Left middle**: an example scenario where four screening instruments identify four different sized groups and they all capture different domains of the construct (e.g., GD-A: mental health problems, GD-B: physical health problems, GD-C: social health problems, GD-D: other problems). **Right middle**: an example scenario where four screening instruments identify four different sized groups of which one involves GRHPs (e.g., GD-1: “gaming disorder” with mixed health problems, other identified groups not related health problems). **On the right**: an example perfect scenario where four screening instruments identify one and the same group. The examples are not exhaustive, and the identified groups need not overlap.

Next, we introduce our hypotheses, which are specified in the Methods section. Based on the recurring findings that show prevalence rates of GRHPs to vary along with the instruments (e.g., Stevens et al. 2020), we expect our ontologically distinct instruments to produce meaningfully different prevalence rates (Appendix 1).

Primary hypothesis, high confidence.

**H1**: *We expect the ICD-11 and DSM-5 based “gaming disorder” prevalence rates to be meaningfully lower than the DSM-IV and Self-assessment[[1]](#footnote-1) based prevalence rates.*

**H1a**: *We expect the ICD-11 based “gaming disorder” prevalence rate to be meaningfully lower than the DSM-IV based prevalence rate.*

**H1b**: *We expect the ICD-11 based “gaming disorder” prevalence rate to be meaningfully lower than the Self-assessment based prevalence rate.*

**H1c**: *We expect the DSM-5 based “gaming disorder” prevalence rate to be meaningfully lower than the DSM-IV based prevalence rate.*

**H1d**: *We expect the DSM-5 based “gaming disorder” prevalence rate to be meaningfully lower than the Self-assessment based prevalence rate.*

The prevalence rate tells us how many people are identified, but it tells little about overlap—*who* are identified. In practice, two identical prevalence rates could identify two entirely different groups of people. For instance, Starcevic and colleagues (2020) assessed 100 individuals who sought treatment for GRHPs and found 36 met both “gaming disorder” and “internet gaming disorder” criteria, whereas additionally 25 met only the latter. Likewise, Ko and colleagues (2020) carried out diagnostic interviews with 69 individuals who met the “internet gaming disorder” criteria and found 44 of them also met the “gaming disorder” criteria (cf. Figure 1). We did not find clinical studies between “game addiction” (DSM-IV pathological gambling based) and the later “internet gaming disorder” (DSM-5) and “gaming disorder” (ICD-11), however, our pilot data indicated that the latter two also meet the DSM-IV based criteria. We hence expect our three related instruments to identify similarly overlapping groups.The overlap of Self-assessment based problems is reported exploratively.

Secondary hypothesis**,** medium confidence.

**H2**: *We expect ICD-11, DSM-5, and DSM-IV based “gaming disorders” to overlap. Those who meet the ICD-11 criteria also meet the DSM-5 criteria, and both above additionally meet the DSM-IV based criteria.*

**H2a**: *Those who meet the ICD-11 criteria also meet the DSM-5 criteria.*

**H2b**: *Those who meet the ICD-11 and DSM-5 criteria also meet the DSM-IV based criteria.*

Finally, we are interested in whether some groups are healthier than others. Because “gaming disorder” is a mental disorder, we expect those identified by related screening tools to have lower mental health scores than the general population. Previous findings also imply similar patterns regarding physical health (Puolitaival et al., 2020). Due to the small prevalence rates acquired in our pilot study (Appendix 1), we expect not to have sufficient power to compare health in ICD-11 and DSM-5 based constructs at Stage 2. Thus, we set a hypothesis on the other two constructs alone. As our pilot gave mixed evidence on the health of those with Self-assessed gaming problems, we set competing H0s.

Tertiary hypothesis, mild confidence.

**H3**: *We expect those with DSM-IV and Self-assessment based “gaming disorders” to have (equally) lower health in comparison to the general population.*

**H3a**: *We expect those with DSM-IV based “gaming disorder” to have meaningfully lower mental or physical health in comparison to the general population.*

**H3b/H0**: *We expect those who Self-assess that they have gaming problems to have meaningfully lower mental or physical health in comparison to the general population.*

**H3c/H0**: *We expect the mental and physical health of those with DSM-IV and Self-assessment based “gaming disorders” to be meaningfully different.*

**Methods**

This study received a positive appraisal from the Human Sciences and Ethics Committee of the University of Jyväskylä in 2021. The research will be carried out according to the Finnish National Board on Research Integrity guidelines and the Helsinki Declaration with its later amendments. A pilot study was carried out to set informed hypotheses (Appendix 1).

***Design***

We will collect a Finnish survey sample (*N*=8000) by using the services of a company specialized in surveys (Bilendi, Appendix 1). The sample will be representative of the Finnish population by gender and region (four regions in Finland) between ages 16 and 65.

**H1**. Among adolescents, GRHPs have been found to be more common than in older groups (e.g., Stevens et al. 2020). Due to the upcoming sample including individuals younger than 18, we expect our adult pilot data prevalence rates to be lower than the ones we will receive next. We have no reason to believe this to notably influence the ratios between the prevalence rates of the four instruments. We expect the differences between the prevalence rates to cohere with those in the pilot (Table 1 in Appendix 1) so that the proportions identified by ICD-11 and the DSM-5 based instruments are meaningfully smaller than those identified by DSM-IV and Self-assessment based instruments (H1a–d). To test the hypotheses, we use an interval-based method, as described in Dienes (2021) (see Neyman & Pearson 1933). We set the H0 critical region (null corroboration) to the lower bound of 95% CI in the smallest obtained prevalence rate and the H1 acceptance region twice above the upper bound of 95% CI in that prevalence rate. The range between these intervals represents an inconclusive region of doubt. If the 95% confidence intervals of the H1a–d effects (differences between prevalence rates) overlap mainly with the H0 or the H1 interval, the hypotheses are corroborated, respectively. If the 95% confidence intervals of the effects (differences between prevalence rates) overlap mainly with the region of doubt, neither H0 nor H1 will be corroborated. H1a–d corroborations would support the position that ICD-11 and DSM-5 based constructs differ from DSM-IV and Self-assessment based gaming problem constructs in terms of prevalence. Due to lacking pilot support for ICD-11 and DSM-5 prevalence differences, we do not set a hypothesis regarding them but report that exploratively. If H0 is corroborated one or more times, this would support the position that the DSM-IV and/or Self-assessment based gaming problem constructs are similar to ICD-11 and/or DSM-5 based ones, prevalence-wise.

**H2**. From previous clinical studies, we found only ICD-11 and DSM-5 based gaming-related health problems analyzed comparatively. These studies found 60% (Starcevic et al. 2020) and 64% (Ko et al. 2020) of “internet gaming disorder” (DSM-5) patients also meeting the “gaming disorder” (ICD-11) criteria, and all the latter meeting all the former. Our pilot found no overlap between the individuals in the ICD-11 and DSM-5 criteria meeting groups. The conflict could be explained by our pilot sample size (Figure A in Appendix 1). It is also known that the content differences between the available screening instruments are large (Karhulahti et al. 2021), thus some of these discrepancies could be explained by the properties of GDT and IGDT10 scales. We thus follow the clinical literature in our hypotheses, in addition to which we expect both the ICD-11 and DSM-5 criteria meeting groups to meet the DSM-IV based cutoff, based on the pilot data: *P*(DSM-5|ICD-11)≈1 and *P*(DSM-5+ICD-11|DSM-IV)≈1. Because mischievousresponding, among similar biases,is a known problem in survey research, we control Type 2 error. In gaming research, Przybylski (2016) found mischievous responding alone to account for up to 2.27% of respondents, indicating that among those who will meet ICD-11 criteria (N=8000x0.2≈16), a false response is not unlikely to occur. Therefore, we allow variation by the lower bound of the binomial probability 50% confidence interval for the conditional probability *P*=1 (*n*=16/16; 0.917, 1). This control allows, for example, one ICD-11 criteria meeting respondent to *not* meet the DSM-5 criteria without undermining the hypothesis (in the example of ICD-11=16; 0.917, 1). If the obtained sample of ICD-11 criteria meeting group meets *P*(ICD-11|DSM-5)>0.917, we consider H2a corroborated and this as evidence for the ICD-11 and DSM-5 based constructs to overlap. For possible null testing, we set the binomial probability confidence interval at 95% for the conditional probability *P*=1 (*n*=16/16; 0.794, 1). If the sample of GDT criteria meeting participants meets *P*(ICD-11|DSM-5)<0.794, we consider H0 corroborated and this as evidence for the ICD-11 and DSM-5 based constructs to be different by overlap. We repeat the process with the same logic regarding ICD-11–DSM-IV and DSM-5–DSM-IV (H2b). Empirical support for overlap can mean that the constructs consist of similar health problems (of different degree), different kinds of health problems, or no health problems (Figure 1). Empirical support for lacking overlap can mean that the constructs are partially or entirely different (Figure 2).

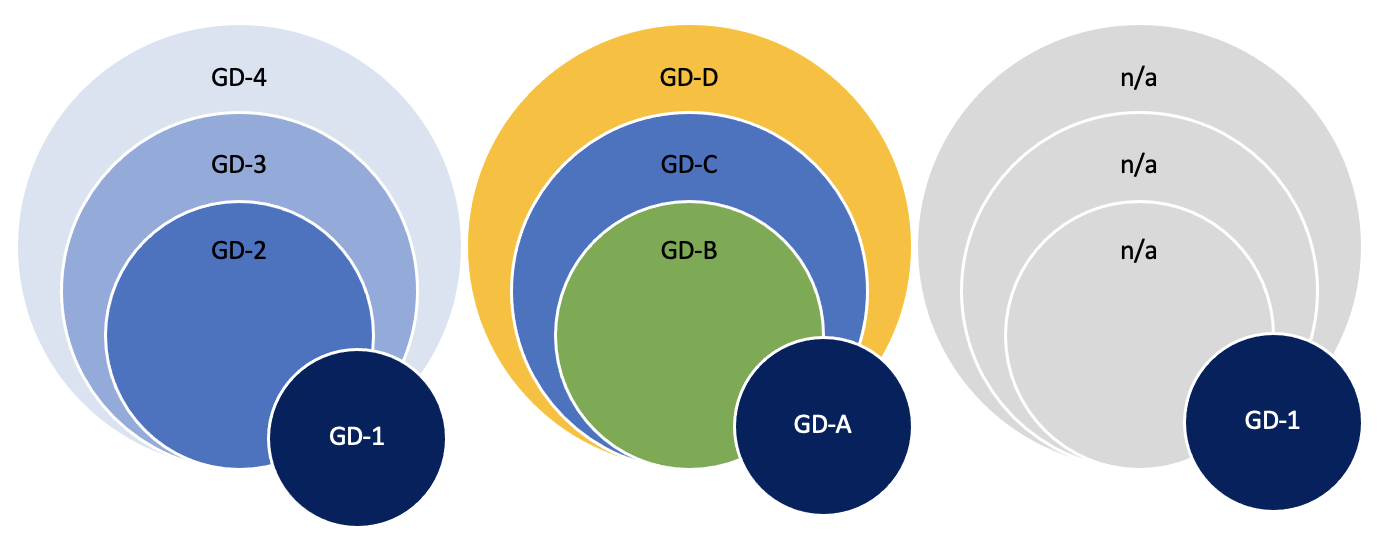


Figure 2. Three exemplary scenarios where one screening instrument identifies a group with marginal overlap.

**H3**. We will apply GPH-2 and GMH-2 (Hays et al. 2017) for investigating general physical and mental health. The United States based reference values for physical health are 42–50 (good) and 35–41 (fair); for mental health, the values are 40–48 (good) and 28–40 (fair) (PROMIS 2021; Hays et al. 2015). When measured with GPH-2, the Finnish working population had the mean value of 44.2 whereas the value for the individuals meeting DSM-IV based criteria had 42.5. With GMH-2, the values were 47.0 for the general working population and 44.5 for the DSM-IV based criteria meeting group. A one-way analysis of variance (ANOVA) test showed that there was a statistically significant difference between both physical health (F[1, 998]=4.66, p<0.05, Cohen’s *d*=0.23, 95% CI [.14, .33]) and mental health (F[1,998]=8.53, p<0.01, Cohen’s *d*=0.32, 95% CI [.22, .41]). Even small effects can be relevant if adverse health scenarios prolong over multiple years (Funder & Ozer 2019); hence, we set our smallest effect size of interest to *d*=0.22 (the lowest 95% CI in the pilot; clear change in reference value). If either the mental or physical health of the DSM-IV and Self-assessment based groups are significantly (*d≥*0.22 at alpha 0.0125) lower than those of the general population, we consider H3a/H3b corroborated. If corroborated, we consider that as theoretical evidence for the Self-assessment or DSM-IV based constructs linking to lower health. If neither mental nor physical effects are significant or below *d*=.22, we consider H3a/H3b not supported and continue with equivalence testing (Lakens 2017) to assess null support. As we did not find strong pilot evidence for H3c (Appendix 1), we set H0 as a competing hypothesis. In H3c, we assess the difference between the DSM-IV and Self-assessment groups.A significant effect (*d≥*0.22 at alpha 0.025) would be interpreted as support for the multiplicity of constructs health-wise. H0 is a competing hypothesis, and its corroboration would be evidence for construct similarity, health-wise.

***Sampling plan***

The data will be collected with Bilendi, i.e. respondents will be invited remotely from the company’s 2.2M panel of participants (Appendix 1). The goal is to recruit 8000 Finnish respondents as nationally representative of gender, region, and age (between 16–65). We include two control questions (Oppenheimer et al. 2009) in the survey and remove those responses that fail both. Additionally, participants who report not having played videogames within the past six months will not fill out GRHP instruments (except THL1 for control purposes), and they will be considered not having GRHPs. The order of the screening instruments will be randomized.

**H1**. Our sample size calculation for the estimated prevalence rates is based on precision rather than power (see Rothman & Greenland, 2018). We calculated the required sample size of 7668 participants based on the 95% confidence interval around the point estimate 0.2, which corresponds with the smallest prevalence rate in the pilot (GDT). Thus, *N*=8000 is a rational sample size for precise prevalence analysis and comparison.

**H2**. The sample size is derived from H1.

**H3**. The sample size requirement for H3a–b is based on power analysis for the Welch t-test (one-sided). For the desired power level of 0.9, alpha set to 0.025 (using multiple comparison correction), and the expected sizes of the groups (GAS7=712; 8.9%; THL1=1104; 13.8%) when the target sample size is *N*=8000, we have power to reliably detect *d*=0.127 and *d*=0.105, respectively. This meets our smallest effect size of interest (*d*=0.22). If either sample will be smaller than the one needed to detect the observed effect, we will not make inferences in H3a–b unless the upper bound of the effect’s confidence interval does not exceed *d*=0.22 (which would support an uninteresting effect). The sample for H3c derives from sensitivity analysis for the Welch t-test (two-sided). Based on the 95% confidence intervals of the above prevalence, sensitivity analysis with the desired power level of 0.9 and alpha set to 0.025 showed that we would be able to reliably detect *d*=0.185 or *d*=0.155, respectively. The code of is available at: <https://osf.io/6fqm5/>

***Analysis plan***

**H1**. We will calculate the prevalence rates with 95% confidence intervals for each of the four instruments. As the ontology of each diagnostic manual is linked to their recommended cutoffs (2nd dimension in Introduction), we follow the recommendations (Appendix 1). We apply the method described by Dienes: “If the CI lies mainly in the H0 interval and the remaining minority only in the grey interval, one could accept H0; similarly, if the CI lies mainly in the H1 interval and the remaining minority only in the grey interval, accept H1; otherwise more data are needed” (2021, p. 9). We set the H0 interval (null corroboration region) to the lower bound of 95% CI in the smallest obtained prevalence rate and the H1 interval (alternative acceptance region) twice above the upper bound of 95% CI in that prevalence rate. Because assessing what “endorsement” of the above items means is not straightforward (Connolly et al. 2021), we repeat the analyses exploratively by using different endorsement criteria.

**H2.** The results regarding cutoff overlap will be presented descriptively based on the above analyses. For H2, we will test group overlap against conditional probability *P*>0.917. The null is tested against conditional probability *P*<0.714.

**H3a–b**. We will compare the means of mental (GMH-2) and physical (GPH-2) health between the DSM-IV based group and the rest of the sample (one-tailed Welch t-test). This is repeated with Self-assessment. Because mental and physical health are measured separately, we carry out the test twice with a corresponding alpha level 0.025 (multiple comparison correction). Equivalence testing is carried out using the TOSTtwo function for an independent t-test (TOSTER package). Exploratively, we repeat the analyses with two general population groups: gaming and non-gaming. We set our equivalence bounds by the same smallest effect size of interest (*d*=0.22). We follow the above in H3c and use THL1 cutoff 2/4 in concordance with GAS7 wording (“sometimes”).

**References**

APA (American Psychiatric Association) 2013. *Diagnostic and Statistical Manual of Mental Disorders*. 5th edition. APA.

Carras, M. C., & Kardefelt-Winther, D. (2018). When addiction symptoms and life problems diverge: A latent class analysis of problematic gaming in a representative multinational sample of European adolescents. *European child & adolescent psychiatry*, *27*(4), 513-525.

Connolly, T., Atherton, G., Cross, L., & Kaye, L. K. (2021). The Wild West of measurement: Exploring problematic technology use cut off scores and their relation to psychosocial and behavioural outcomes in adolescence. *Computers in Human Behavior, 125*, e106965.

Dienes, Z. (2020). Obtaining evidence for no effect. *Collabra: Psychology* (in press). Preprint: <https://doi>.org/10.31234/osf.io/yc7s5

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, 2(2), 156-168.

Hays, R. D., Spritzer, K. L., Thompson, W. W., & Cella, D. (2015). US general population estimate for “excellent” to “poor” self-rated health item. *Journal of General Internal Medicine*, 30(10), 1511-1516.

Hays, R. D., Schalet, B. D., Spritzer, K. L., & Cella, D. (2017). Two-item PROMIS® global physical and mental health scales. Journal of patient-reported outcomes, 1(1), 1-5.

Karhulahti, V., Martončik, M., & Adamkovic, M. (2021). Measuring Internet Gaming Disorder and Gaming Disorder: Qualitative Content Validity Analysis of Validated Scales. *Assessment*.

King, D. L., Chamberlain, S. R., Carragher, N., Billieux, J., Stein, D., Mueller, K., … Delfabbro, P. H. (2020). Screening and assessment tools for gaming disorder: A comprehensive systematic review. Clinical Psychology Review, 77, 101831. Doi:10.1016/j.cpr.2020.101831

Király, O., Sleczka, P., Pontes, H. M., Urbán, R., Griffiths, M. D., & Demetrovics, Z. (2017). Validation of the Ten-Item Internet Gaming Disorder Test (IGDT-10) and evaluation of the nine DSM-5 internet gaming disorder criteria. Addictive Behaviors, 64, 253–260.

Ko, C. H., Lin, H. C., Lin, P. C., & Yen, J. Y. (2020). Validity, functional impairment and complications related to Internet gaming disorder in the DSM-5 and gaming disorder in the ICD-11. Australian & New Zealand Journal of Psychiatry, 54(7), 707-718.

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science*, *8*(4), 355-362.

Lemmens, J. S., Valkenburg, P. M., & Peter, J. (2009). Development and validation of a game addiction scale for adolescents. Media psychology, 12(1), 77-95.

Myrseth, H., & Notelaers, G. (2018). A latent class approach for classifying the problem and disordered gamers in a group of adolescence. *Frontiers in psychology*, *9*, 2273.

Neyman, J. & Pearson, E. (1933). The testing of statistical hypotheses in relation to probabilities a priori. *Mathematical Proceedings of the Cambridge Philosophical Society*, 29(04), 492–510.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of experimental social psychology*, *45*(4), 867-872.

Pontes, H. M., Schivinski, B., Sindermann, C., Li, M., Becker, B., Zhou, M., & Montag, C. (2019). Measurement and conceptualization of gaming disorder according to the World Health Organization framework: The development of the Gaming Disorder Test. *International Journal of Mental Health and Addiction*.

PROMIS (2021) PROMIS® Reference Populations. Official instrument website. <https://www.healthmeasures.net/score-and-interpret/interpret-scores/promis/reference-populations>

Przybylski, A. K. (2016). Mischievous responding in Internet gaming disorder research. *PeerJ*, *4*, e2401.

Puolitaival, T., Sieppi, M., Pyky, R., Enwald, H., Korpelainen, R., & Nurkkala, M. (2020). Health behaviours associated with video gaming in adolescent men: a cross-sectional population-based MOPO study. BMC public health, 20(1), 1-8.

Rothman, K. J., & Greenland, S. (2018). Planning Study Size Based on Precision Rather Than Power. Epidemiology, 29(5), 599–603.

Salonen, A. & Raisamo, S. (2015) *Suomalaisten rahapelaaminen 2015: Rahapelaaminen, rahapeliongelmat ja rahapelaamiseen liittyvät asenteet ja mielipiteet 15-74-vuotiailla*. Helsinki: THL.

Starcevic, V., Choi, T. Y., Kim, T. H., Yoo, S. K., Bae, S., Choi, B. S., & Han, D. H. (2020). Internet gaming disorder and gaming disorder in the context of seeking and not seeking treatment for video-gaming. Journal of psychiatric research, 129, 31-39.

Stevens, M. W., Dorstyn, D., Delfabbro, P. H., & King, D. L. (2020). Global prevalence of gaming disorder: A systematic review and meta-analysis. Australian & New Zealand Journal of Psychiatry, 0004867420962851.

THL (2021) National PROMIS Centre. Finnish Institute for Health and Welfare. <https://thl.fi/fi/web/toimintakyky/toimintakyvyn-arviointi/kansallinen-promis-keskus>

**Acknowledgements**

The work was supported by the Finnish Work Environment Fund (200349), Academy of Finland (312397), and the Slovak Research and Development Agency under the Contract no. APVV-18-0140.

**Author contributions**

Conceptualization: VMK, JV, RK, MB. Data curation: JV, VMK. Formal analysis: JV, MMA, MMU, MB. Funding Acquisition: VMK, JV, MMU, RK, MB. Investigation: VMK, JV, MMU, RK, MB. Methodology: VMK, JV, MMA, MMU, MB. Project administration: VMK. Resources: VMK, RK. Software: JV, MMA, MMU, MB. Supervision: N/A. Validation: VMK, MMU, MB. Visualization: VMK. Writing original draft: VMK. Writing review & editing: VMK, JV, MMA, MMU, RK, MB.

**Competing interests**

VMK is one of the PCI Registered Reports recommenders. JV is a CSO in Kinrate Analytics, which provides personalized game recommendations.

**Appendix 1**

To be able to set informed expectations for our hypotheses, we acquired pilot data (*N*=1000) by using the recruitment services of the company Bilendi (a company specialized in recruiting survey samples). Data collection took three days—the company estimated that our upcoming *N*=8000 might take up to a month to collect—and the participants were incentivized by the company with soft currency points, which can be transformed into material reward items. The participants filled out the survey remotely by using digital devices (computer, tablet, smartphone). The respondents were Finnish adults, and the pilot sample was nationally representative of the working population of Finland by age (*M*=41.6, *SD*=12.2), region (equal distribution between four Finnish regions) and gender (men *n*=521). The respondents filled in the previously described four screening instruments (also others, see later). The Finnish translations of the first two instruments have been found reliable in previous studies (GAS7: Männikkö et al. 2015; IGDT10: Männikkö et al. 2019), and the Self-assessed gaming problems instrument (THL1) was originally developed in Finnish language by the ministry operated Finnish Institute for Health and Welfare. Because GDT is a new instrument and does not have validation studies of its translated versions yet, we (back) translated it with the help of an English-Finnish linguist and carried out convergent validation tests for the translated version (Supplement). The translated GDT had composite reliability (CR) of 0.90 and average variance extracted (AVE) value of 0.69, both of which exceeded their required levels (CR>0.70, AVE >0.50). Hence, the tests supported convergent validity for the translated GDT (see Fornell and Larcker, 1981).

For GDT, which is a 4-item instrument with a five-point scale ( “Never,” “Rarely,” “Sometimes,” “Often,” “Very Often”), we use a monothetic cutoff to identify those who answer “Often” or “Very Often” to all items (cutoff=4/4 according to ICD ontology). For IGDT10, which is a 10-item instrument with a three-point scale (“Never,” “Sometimes,” “Often”), we use a polythetic cutoff to identify those who answer “Often” to at least five items (the last two items are combined to give one point, cutoff=5/9 according to DSM ontology). For GAS7, which is a 7-item instrument with a five-point scale (“Never,” “Rarely,” Sometimes,” “Often,” “Very Often”), we use a polythetic cutoff to identify those who answer “Sometimes” to at least four items (cutoff=4/7 according to DSM ontology). For THL1, which is a 1-item instrument with a four-point scale (from “Never,” Sometimes,” Often,” “Almost Always”), we report both previously applied (Karhulahti & Koskimaa 2019; Salonen & Raisamo 2015) cutoffs to identify those who self-assess problems at least “Sometimes” and “Often” (cutoff=1/1). The prevalence rates of GRHPs according to each instrument are presented in Table 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ONTOLOGICAL BASIS** | **ICD-11**  gaming disorder | **DSM-5**  internet gaming disorder | **DSM-IV**  pathological gambling | **Self-assessed** |
| **instrument** | GDT | IGDT10 | GAS7 | THL1 |
| **prevalence** | **0.2 %**, *n*=2  95% CI (0.05–0.79) | **0.6 %**, *n*=6  95% CI (0.24–1.33) | **8.9 %**, *n*=89  95% CI (7.28–10.83) | * 1. **%**, *n*=24   95% CI (1.61–3.56) |

**Table 1**. Prevalence rates in the pilot. These rates were calculated with Stata/SE 16.1. We did a verifying analysis with IBM SPSS Statistics 26, which was identical except for GAS7 having 91 participants in the group. After additional calculations with R that produced the original results, we concluded this higher prevalence for GAS7 in the SPSS being associated with a software flaw. Other analyses were also replicated with SPSS (no discrepancies). Note: the explorative THL1 is presented here by its higher cutoff (3/4), while the lower cutoff (2/4) had a prevalence of *n*=138 (13.8; 95% CI 11.80–16.07).

Related to our H2, we analyzed the overlap between the above four groups. Even though almost all individuals who met the cutoff by GDT and IGDT10 also met the cutoff by GAS7, none of those who met the GDT cutoff met the IGDT10 cutoff. The overlap is presented in Figure A.

Diagram

Description automatically generatedDiagram

Description automatically generated

Diagram

Description automatically generatedDiagram

Description automatically generated

**Figure A**. The groups identified by each of the four instruments in the pilot. We present only the higher cutoff (3/4) for THL1.

For H3, the data provided us with estimates for mental and physical health differences. The respondents filled out GPH-2 and GMH-2 scales (Hays et al. 2017), which have been localized and tested by the collaboration of the National Institute for Health and Welfare and PROMIS Health Organization (THL 2021). Because the Finnish translation does not have published validation studies yet, we tested their internal consistencies (existing Stata code <https://www>.stata.com/manuals13/mvalpha.pdf). The coefficient alpha for GMH-2 was good (0.78) and for GPH-2 slightly below (0.66). Spearman-Brown produced the same results. Due to the instruments having only two items, and the translations having been done carefully with a professional linguist, we considered the instruments valid for use in this study.

The sample size allowed us to compare the GAS7 criteria meeting group (*n*=89) mental and physical health with the general population. In terms of both mental and physical health, the general population had slightly higher mean values than the individuals identified by GAS7 (see reference values in the Methods section). A one-way analysis of variance (existing State code <https://www>.stata.com/manuals13/mvalpha.pdf) test showed that there was a statistically significant difference between both physical health [F(1, 998)=4.66, *p*<0.05, Cohen’s *d*=0.23 (95% CI .14, .33)] and mental health [F(1,998)=8.53, p<0.01, Cohen’s *d*=0.32 (95% CI .22, .41)]. When identifying a group by the lower THL1 cutoff 2/4 (*n*=138), the effects were smaller [lower physical health (*p*=0.09, *d*=-0.13, 95% CI [-.30, .05]), lower mental health (*p*=0.04, *d*=-0.16, 95% CI [-.34, .02)] and the CIs crossed both null and our smallest effect size of interest. To seek clarifying support, we accessed and analyzed a recently opened nationally representative dataset (*N*=3994), shared by the Finnish Health Institute (THL & Salonen 2021). The data were collected for gambling research, but it also included modified GAS7 (timespan changed from 6 months to 12 months and “gaming” replaced by “digital gaming”), the original THL1, and Mental Health Inventory 5 (Berwick et al. 1991). We found both GAS7 (*p*=.015, *d*=0.37, 95% CI .14–2.26) and THL1 (cutoff 2/4; *p*=.000, *d*=0.52, 95% CI .35–.65) to correlate with lower mental health scores. This supported our H3a, but provided mixed evidence for H3b/c. Therefore, we decided to set competing H0 hypotheses for the latter.

We use these pilot data to specify outcome expectations for our hypotheses in the analysis plan section. Materials are available here: <https://osf.io/6fqm5/>

The final survey will also measure gambling problems (Salonen & Raisamo 2015; translation from Gebeuer et al. 2010), perceived loneliness (Tilvis et al. 2000), work ability (WAI short version; Ilmarinen 2007), and work recovery (Mauno et al. 2017), which we will use exploratively. The results will not be reported in this registered report. Table 2 summarizes all measures.

|  |  |  |
| --- | --- | --- |
| **Construct** | **Measure** | **Details** |
| *Gaming-related health problems* | | |
| Gaming addiction | Gaming addiction scale 7 (GAS7), Lemmens et al. (2009). | 7-item measure on scale 1–5, polythetic cutoff 4/7 by score 3/5 |
| Internet gaming disorder | Internet Gaming Disorder Test 10 (IGDT10), Király et al. (2017). | 10-item measure on scale 1–3, polythetic cutoff 5/9 by score 3/3 |
| Gaming disorder | Gaming disorder test (GDT), Pontes et al. (2019). | 4-item measure on scale 1–5, monothetic cutoff 4/4 by score 4/5 |
| Self-assessed problem gaming | Problem gaming test (THL1), Salonen & Raisamo (2015). | 1-item measure on scale 1–4, cutoff by score ¾ |
| *General health* | | |
| General Mental Health | PROMIS General Mental Health (GMH-2), Hays et al. (2017). | In US national standards, less than 28 = poor, more than 56 = excellent |
| General Physical Health | PROMIS General Physical Health (GPH-2), Hays et al. (2017) | In US national standards, less than 35 = poor, more than 58 = excellent |
| *Additional measures (not used in this registered report)* | | |
| Work ability | Work ability index short form (WAI), Ilmarinen (2007) | 1-item measure on scale 0–10 |
| Work recovery | Work recovery question (WR), Mauno et al. (2017) | 1-item measure on scale 1–5 |
| Perceived loneliness | Perceived loneliness (PL), Tilvis et al. (2000) | 1-item measure on scale 1–3 |
| Pathological gambling | BBDS, Salonen & Raisamo (2015) from Gebauer et al. (2010) | 3-item measure on scale 1–2, polythetic cutoff 1/3 by score 2/2 |

**Table 2**. All measures.

***References***

Berwick, D. M., Murphy, J. M., Goldman, P. A., Ware Jr, J. E., Barsky, A. J., & Weinstein, M. C. (1991). Per- formance of a five-item mental health screening test. Medical care, 169–176.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. Journal of Marketing Research, 18(1), 39–50. <https://doi.org/10.2307/3151312>.

Gebauer L, LaBrie R, Shaffer HJ. Optimizing DSM-IV-TR Classification accuracy: A brief Biosocial Screen for Detecting Current gambling Disorders Among Gamblers in the General Household Population. The Canadian Journal of Psychiatry 2010; 55(2): 82-90.

Hays, R. D., Schalet, B. D., Spritzer, K. L., & Cella, D. (2017). Two-item PROMIS® global physical and mental health scales. Journal of patient-reported outcomes, 1(1), 1-5.

Ilmarinen, J. (2007). The work ability index (WAI). *Occupational medicine*, *57*(2), 160-160.

Karhulahti, V. M., & Koskimaa, R. (2019). On the Prevalence of Addicted or Problematic Gaming in Finland. Addictive behaviors reports, 10, 100225.

Mauno, S., Hirvonen, R., & Kiuru, N. (2018). Children’s life satisfaction: The roles of mothers’ work engagement and recovery from work. *Journal of Happiness Studies*, *19*(5), 1373-1393.

Männikkö, N., Billieux, J., & Kääriäinen, M. (2015). Problematic digital gaming behavior and its relation to the psychological, social and physical health of Finnish adolescents and young adults. Journal of behavioral addictions, 4(4), 281-288.

Männikkö, N., Ruotsalainen, H., Tolvanen, A., & Kääriäinen, M. (2019). Psychometric properties of the Internet Gaming Disorder Test (IGDT‐10) and problematic gaming behavior among Finnish vocational school students. Scandinavian journal of psychology, 60(3), 252-260.

Salonen, A. & Raisamo, S. (2015) *Suomalaisten rahapelaaminen 2015: Rahapelaaminen, rahapeliongelmat ja rahapelaamiseen liittyvät asenteet ja mielipiteet 15-74-vuotiailla*. Helsinki: THL.

THL (2021) National PROMIS Centre. Finnish Institute for Health and Welfare. <https://thl.fi/fi/web/toimintakyky/toimintakyvyn-arviointi/kansallinen-promis-keskus>

THL & Salonen, Anne (National Health Institute): Rahapelitutkimus 2019 [electronic data]. Version 1.0 (2021-04-27). Finnish Social Science Data Archive. <http://urn>.fi/urn:nbn:fi:fsd:T- FSD3522

Tilvis, R., Pitkälä, K., Jolkkonen, J., & Strandberg, T. (2000). Feelings of loneliness and 10-year cognitive decline in the aged population. *Lancet*, 356, 77-78.

**Appendix 2**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **RQ** | **H** | **Sampling** | **Analysis** | **Rationale** | **Interpretation** | **Theory** |
| **A** | **1a**  **1b**  **1c**  **1d** | Nationally representative *N*=8000, which corresponds with the precision required to assess prevalence for 0.2 (*n*=7668). This prevalence is the lowest one found in our pilot (for GDT). | We calculate prevalence rates and 95% confidence intervals with recommended cutoffs for the screening instruments. We apply the method described by Dienes: “If the CI lies mainly in the H0 interval and the remaining minority only in the grey interval, one could accept H0; similarly, if the CI lies mainly in the H1 interval and the remaining minority only in the grey interval, accept H1; otherwise more data are needed” (2021, p. 9). | All instruments have a different ontological basis. Significant differences between prevalence rate proportions would be some evidence for these ontological bases representing different constructs. No difference would be evidence for these ontological bases representing the same construct. | In each hypotheses, we set the H0 interval to the lower bound of the smallest obtained prevalence rate and the H1 interval to x2 of its upper bound. If the confidence intervals of the *differences between compared prevalence rates* fall mainly to the H1 area, we consider H1a–d corroborated. Null will be corroborated by the confidence intervals falling mainly to the H0 area. If the confidence interval falls mainly in the grey area between, none of the hypotheses are corroborated. | H1 (if all sub hypotheses are corroborated) would support the position that the DSM-IV and/or Self-assessment based “gaming disorder” constructs are different from those of ICD-11 and/or DSM-5, prevalence-wise. H0 (if all sub hypotheses are refuted) would support the similarity of the constructs, respectively. Both types of evidence would be theoretically meaningful. In case of mixed findings, we do not draw full theoretical implications for H1 or H0. |
| **B** | **2a** | N/A | We test *P*(DSM-5|ICD-11)≈1 and *P*(DSM-5+GDT|DSM-IV)≈1. To control Type 2 error, we allow variation by the lower bound of the binomial probability 50% confidence interval for the conditional probability *P*=1 (at n=16/16; 0.917, 1). For possible null testing, we set the binomial probability confidence interval at 95% for the conditional probability *P*=1 (n=16/16; 0.794, 1). | All instruments have a different ontological basis, but previous literature has shown 100% of those who meet ICD-11 based criteria to also meet DSM-5 criteria. We thus expect the ICD-11 criteria meeting group to also meet the DSM-5 cutoff. To control Type 2 error, we allow variation based on the 50% confidence interval of *P*=1 (at n=16/16; 0.917, 1). | If the obtained sample of ICD-11 identified participants meets *P*(ICD-11|DSM-5)>0.917, we consider H2a corroborated and this as evidence for the ICD-11 and DSM-5 based constructs to be similar by overlap. If the obtained sample of GDT-identified participants meets *P*(ICD-11|DSM-5)<0.794,  we consider this as evidence for the ICD-11 and DSM-5 based constructs to be different by overlap. | If both H2a and H2b are corroborated, this would be evidence for all ICD-11, DSM-5, and DSM-IV based constructs to overlap. If H2a and H2b are not corroborated, this will be evidence for construct differences (three possible differences: ICD-11 vs DSM-5, ICD-11 vs DSM-IV, DSM-5 vs DSM-IV). Similarities support the singular “gaming disorder” construct and differences support the multiplicity thereof. In case of mixed findings, we do not draw full theoretical implications for H2 or H0. |
| **B** | **2b** | Our pilot indicates both ICD-11 and DSM-5 based groups to also meet DSM-IV criteria. We thus expect the ICD-11 and DSM-5 criteria meeting groups to also belong to the DSM-IV criteria meeting group (in the same way as above). | [Same as above, with GDT, IGDT10, GAS7] |
| **C** | **3a** | The sample size is based on a power analysis for the Welch t-test (one-sided). We expect that the size of the effect between the DSM-IV group and the rest of the sample will be approximately *d*=0.22–0.41 with GPH-2 and GMH-2. For the desired power level of 0.9, alpha set to 0.0125, and the expected sizes of the DSM-IV group (*n*=712 / 8.9%) and self-assessment group (*n*=192 / 2.4%) when the target sample size is *N*=8000, we have power to reliably detect *d*=0.138 and *d*=0.257, respectively. Although the latter does not meet our lowest effect size of interest (*d*=0.22), we have reason to believe that our prevalence rates will be higher. If the self-assessment criteria meeting group will be smaller than the one needed to detect the observed effect, we will not make inferences in H3b unless the upper bound of the effect’s confidence interval does not exceeds our smallest effect size of interest (which would support uninteresting effect). | We compare the means of mental and physical health, measured by GPH-2 and GMH-2, between the DSM\_IV group and the rest of the sample (two-tailed ANOVA). This is repeated with self-assessment criteria. We carry out the Welch t-test twice with a corresponding alpha level 0.0125 (multiple comparison correction). In case of nonsignificant results, we continue with equivalence testing (as below). | Based on pilot data, we expect the DSM-IV criteria meeting group to have lower general mental health than the remaining sample. Previous evidence and our pilot indicate the same for physical health, and we expect this, too. | If either the mental or physical health of the DSM-IV based criteria meeting groups are significantly (d*≥*0.22 at alpha 0.0125) lower than those of the rest of the sample, we consider H3a corroborated. A nonsignificant or below *d*=0.22 effect will not corroborate the H3a. | If corroborated, we consider that as evidence for the THL1 (self-assessed) or GAS7 (DSM-IV pathological gambling based) criteria meeting groups to have lowered mental and/or physical health (no causal interpretations). This would support the two instruments to measure one or more constructs, which are distinct health-wise in relation to the general population. Equivalence, in turn, would support equal health levels between those with “gaming disorder” in the two respective ontological domains and the general population. |
| **C** | **3b** | [Same as above with mixed evidence for THL1] | Same as above with THL1, where equivalence testing can support the null.] |
| **C** | **3c** | N/A | We use the TOSTER::TOSTone.raw() function from the TOSTER package. First, we will compute raw means (mdiff1, mdiff2) and standard deviations (sddiff1, sddiff2) for both effects (differences between groups based on the GAS7 and THL1 classification). In TOSTER::TOSTone.raw() function *m* = 0, *sd =* (sddiff1, sddiff2)/2 and *mu* = mdiff1 – mdiff2. Lower and upper equivalence bound will be +- 0.2 *sd*. If we do not find equivalence, we move to test significance both ways. | Based on pilot data, both equivalence and difference are possible. We set two competing hypotheses. | We test significance between the health levels of the two groups and consider d*≥*0.22 (at alpha 0.025) evidence for difference. | Corroborated H0 would weakly support self-assessed “gaming disorder” and DSM-IV based “gaming disorder” to represent the same or similar construct health-wise. If we find a difference, this would some evidence for the multiplicity regarding the two constructs health-wise. Fully corroborated H3/H0 would require all sub hypotheses to be corroborated. |

**RQ**: How do screening instruments that derive from different ontological understandings differ in identifying GRHP groups?

**RQ-A**: How do GRHP screening instruments that derive from separate ontological understandings differ in *their prevalence rates (how many)*?

**RQ-B**: How do GRHP screening instruments that derive from separate ontological understandings differ in *who they identify (what characteristics)*?

**RQ-C**: How do GRHP screening instruments that derive from separate ontological understandings differ inthe *health of their identified groups (how healthy)*?

**H1**: *We expect the ICD-11 and DSM-5 based “gaming disorder” prevalence rates to be meaningfully lower than the DSM-IV and Self-assessment based prevalence rates.*

**H1a**: *We expect the ICD-11 based “gaming disorder” prevalence rate to be meaningfully lower than the DSM-IV based prevalence rate.*

**H1b**: *We expect the ICD-11 based “gaming disorder” prevalence rate to be meaningfully lower than the Self-assessment based prevalence rate.*

**H1c**: *We expect the DSM-5 based “gaming disorder” prevalence rate to be meaningfully lower than the DSM-IV based prevalence rate.*

**H1d**: *We expect the DSM-5 based “gaming disorder” prevalence rate to be meaningfully lower than the Self-assessment based prevalence rate.*

**H2**: *We expect ICD-11, DSM-5, and DSM-IV, based “gaming disorders” to overlap. Those who meet the ICD-11 criteria also meet the DSM-5 criteria, and both above additionally meet the DSM-IV based criteria.*

**H2a**: *Those who meet the ICD-11 criteria also meet the DSM-5 criteria.*

**H2b**: *Those who meet the ICD-11 and DSM-5 criteria also meet the DSM-IV based criteria.*

**H3**: *We expect those with DSM-IV and Self-assessment based “gaming disorders” to have (equally) lower health in comparison to the general population.*

**H3a**: *We expect those with DSM-IV based “gaming disorder” to have meaningufully lower mental or physical health in comparison to the general population.*

**H3b/H0**: *We expect those who Self-assess that they have gaming problems to have meaningfully lower mental or physical health in comparison to the general population.*

**H3c/H0**: *We expect the mental and physical health of those with DSM-IV and Self-assessment based “gaming disorders” to be meaningfully different.*

1. All the applied instruments are self-assessment survey tools, but only one (THL1) asks the respondent to directly assess their gaming problems. [↑](#footnote-ref-1)