Data Analysis Plan

Psychological Distress in Emerging Adulthood: A Longitudinal Study

DistressEA
ClinicalTrials.gov Identifier: NCT04596345

Date: August 6 ,2021
Updated: February 28, 2022

Data Analysis Plan (DAP) for the trajectories of success and/or distress study

Contents

History/evolution of the project and DAP 1
Introduction ..... **Error! Bookmark not defined.**
PPDAC approach 3
References 11

History/evolution of the project and DAP

- The design of the project was first registered, 22/10/20, before data collection started: https://clinicaltrials.gov/ct2/show/NCT04596345.
- Following that registration, but after data collection had started, before any data analysis has started and prior to submission of a protocol paper for the study the first shared iteration of this DAP was signed and dated (6/8/21).
- Subsequent signed, dated and publicly registered amendments to the DAP will be made when indicated by:
    - Recognition of errors in an earlier iteration: these will be marked as such.
    - Learning from the baseline analysis revealing the need for an amendment but prior to anay analysis of the data affected. The changes will be clearly marked and the rationale for them given.
    - Changes in analyses or new analyses required by peer reviewers (where we accept the arguments for them). These again will be marked and the rationale given.
    - Recognition of emerging issues in the data requiring analytic methods of sufficient complexity to justify documentation. These may be marked as new DAP subsections.
    - Discovery of new methods that might or clearly do promise advantages over the original planned analyses. Again these will clearly marked as such and may form new DAP subsections.
- This iteration, v.2.2 is dated 28/2/22 after submission of the paper analysing the baseline data and immediately before termination of all data collection and before any inspection of the change data. Changes were in readability and correction of typos, not substantive changes in any planned analyses.

Epistemological and methodological position

- Our epistemological position is pragmatic: the data and derived findings should ultimately inform thinking about what interventions may be helpful to optimise mental health of this cohort and their status and legitimacy as evidence is always to be evaluated in that frame being aware of the contextual epistemologies of the likely audiences and dominant methodological expectations.

- In this epistemological position this study is best regarded as a well powered but exploratory test of the methods contributing a substantial, but tentative, body of new knowledge.
- The data are largely quantitative and the approach largely statistical within a statistical tradition of (Abelson, 1995; Spiegelhalter, 2019). We also follow the principles of (Wagenmakers et al., 2021, 'Seven steps toward more transparency in statistical practice') specifically:

  (1) visualizing data;
  (2) quantifying inferential uncertainty;
  (3) assessing data preprocessing choices;
  (4) reporting multiple models;
  (6) interpreting results modestly; and
  (7) sharing data and code.

  Their fifth step, involving multiple analysts is beyond our resources but we welcome offers from others to reanalyse our data, ideally in collaboration but of course our making the data and code available publicly makes it possible for anyone to reanalyse our data. Particular reference to these steps are noted against particular aspects of our analyses below.
  - Some elements are frankly exploratory and pathfinding: mainly the quantitative and qualitative data from the "What's Going On?" (WGO, see registration and protocol) and their relationship with other variables. We are particularly careful to present these as such and not to overclaim for them (Wagenmakers et al., 2021, steps 1, 2, 4 and 6).
  - In general the statistical position is descriptive and based on estimation rather than formal null hypothesis significance testing (NHST) and we use confidence intervals (CIs) to understand the probable impacts of sample size on precision of estimation of population parameters. Where statistical tests are used, they aim to give a pragmatic indication that non-random effects exist, not to give strong assertions that "x is a statistically significant effect" but, with the CIs, to help disentangle seemingly systematic effects from random ones. (Wagenmakers et al., 2021, steps 2, and 6)
  - We recognise that the sample is not a random one so caution will be exercised in any generalisations to the wider population (Wagenmakers et al., 2021, steps 2 and 6).
  - We recognise that there will be incomplete responding and that a small proportion of participants may cross from one baseline group, i.e. students or non-students, to the other group. The impacts of these issues will be explored (Wagenmakers et al., 2021, steps 2, 3, 4 and 6).
  - As in all survey data variables will not be randomly related to one another but may show statistically significantly associations complicating analyses explored, this will be noted and explored carefully (Wagenmakers et al., 2021, steps 2, 3, 4 and 6).
  - We believe that the self-report measures should always be treated not as if they were "hard measures" (Paz, Adana-Díaz, et al., 2020) but as what they are: soft but available and useful measures. Accordingly, reflective exploration of the psychometric properties of the measures will form a part of the study, particularly for the novel "What's Going On?" measure which has only been used in one study before (still awaiting formal publication but available at

https://pure.roehampton.ac.uk/ws/portalfiles/portal/6879104/Blackshaw_Emily_Final_Thesis.pdf).

- This DAP pre-specifies a set of "*a priori*" analyses to optimise likely reproducibility and minimise bias and the huge impacts of *post hoc* analyses on coverage CIs and bias and coverage of p values. However, it is not intended to pre-empt appropriate analyses of unexpected, emergent findings. Equally it is accepted that there may be analytic methods we have not identified at this point which may later be found and be clearly appropriate to add value to the data. All such analyses will be clearly described as such and all publications will refer to a dated, publicly accessible iteration of this document finalised before the analyses described here were started to enable readers to confirm this distinction between *a priori* analyses and *post hoc* analyses of emergent findings.

- The work has been designed using the PPDAC: Problem, Plan, Data, Analyses, Communication approach (Abelson, 1995; Oldford & MacKay, 2000).

Ethics

Formal ethical approval is covered in the registration. More generally we follow the requirements of (Altman & Moher, 2013) but remain careful about the dangers of simply relying on institutional ethical structures (Åm, 2018) and, while our data and methodology are largely quantitative, as a transcultural, cross-language collaboration we find the principles of openness, reflexivity and pragmatism as described in (Erhard et al., 2021) helpful and in line with our principles of collaboration that acknowledges the power of culture, location, language, politics and power.

PPDAC approach

- P ("Problem" in PPDAC terminology is approximately the same as the primary aim of the work.)
  To understand more of the levels of distress of emerging adults in Ecuador over a one year period.
  - Specific questions:
    i. What, for the purposes of comparisons between people, do the baseline data say about the psychometric quality of the two multi-item measures (CORE-OM, EQ-5D-3L; see registration)?
    ii. What do the change data say about the psychometric quality of the two multi-item measures as measures of change and particularly of intra-individual change?
    This is not the same issue as (i) though traditional psychometric methods tend to be presented as if it were the same so it may be wise to expand this distinction here.
    Traditional psychometric methods are designed to address differences between individuals and when considered to address our next question (iii below) such methods are appropriate. However, in order to assess the psychometric qualities of the measures *for measurement of within-participant change* traditional methods must axiomatically assume that individual change has the same covariance structure as the covariance structure of score comparisons between individuals.
    In fact there is no reason to believe that this will be true for complex psychological variables like well-being or quality of life. For such variables it

is more plausible that there are within-participant patterns of change across multiple items just as it is plausible that temporal variation is not the same for all individuals. See (Molenaar, 2004) for a trenchant exposition of the issues.

For example it may be that some participants have fluctuating problems with social functioning while others, say, struggle with confidence but not social functioning. If this is true over the year, then the covariance structure of change across CORE-OM items within the person struggling with social functioning would be different from that of the person struggling with confidence and neither need show an item change covariance matrix resembling the item score covariance matrix across all participants at baseline. However, as long as sufficient items are affected for each of those participants, the CORE-OM-NR would still be a useful change measure for each participant and that utility would be unrelated to the baseline conventional reliability of the measure.

This raises the question whether there are simple analyses that throw more appropriate light on the psychometrics of using the questionnaires to measure individual change than do the conventional explorations of psychometric quality of multi-item questionnaire measures for cross-sectional comparisons between people. This will be explored (details in the PPDAC "A" section: analyses, below).

iii. As most existing data on the well-being and mental health problems of this age group are from student samples, a primary focus will be on differences between a student and a non-student sample both at baseline and subsequently through the year.

iv. How much do scores on the dependent measures (CORE-OM, EQ-5D-3L and EQ-VAS, see registration) vary across the seven time points and is there evidence that the level of variation varies across the sample, i.e. that temporal stability is not a consistent parameter but show strong individual differences? Analyses of the CORE-OM will consider the non-risk and risk scores only (but see “supplementary analyses”).

v. Does baseline group, student vs. non-student, appear to relate to the dependent variables? Do other demographic predictors relate to the dependent variables? Recognising that “the dependent variables” will have temporal variation this will be divided into two sets of analyses:
    1. Using baseline scores only
    2. Analysing the full change data (more details below):
        a. Testing for difference in any systematic linear trend over the seven time points.
        b. Against within-participant SD of the scores.
        c. Against highest and lowest within-participant difference from baseline.

vi. What levels of and changes in stressors/life events (LE) do participants report on the novel “What’s going on?” (WGO) measure?

vii. Is there evidence of systematic relationship between stress and dependent variable scores within-participant over time?

viii. Given the inevitably non-random nature of the sampling creating the dataset this will be compared with the most recently available census data

from Quito for this age group to see how much the sample differs from the census data where comparable variables are present in the census data.

ix. Scores on the CORE-OM will also be compared with those from the only two existing Ecuadorean datasets of student and non-student participants (Paz, Mascialino, et al., 2020) while recognising that as those data also arose from non-random sampling they don't have the same referential status as the comparisons with the census data.

- **Supplementary analyses**. The following will involve combining item data from the baseline of this study with other Ecuadorean CORE-OM data, non-help-seeking and help-seeking, already collected by the research team.
  - Domain and short form scoring.
    As well as the total score across all 34 items the original design of the CORE-OM (Evans et al., 2000, 2002) described five domain scores while recognising that these were not expected to form clean factors in any empirical datasets. These domains were: well-being (W), problems (P), functioning (F) and risk (R). However, a total score across the three non-risk domains to create a non-risk score (NR) has often been encouraged to be analysed in parallel with this R score, reducing the measure to those correlated but relatively empirically and conceptually distinct scores.
    In addition, four shortened forms derived from the CORE-OM are well established (Evans, 2012); these are two 18 item short forms (with the same domains and NR scoring), a 14 item GP-CORE for general population survey work and the CORE-10 for repeated use with help-seeking or otherwise at-risk populations. Though this is not a primary objective of the study, baseline distributions and psychometric exploration of all these scores from the items embedded in the CORE-OM will be reported as these will add importantly to the emerging referential and psychometric data for use of CORE measures in Ecuador and across Latin America.
  - CORE-6D scoring.
    In addition to the scores and the short form scores noted immediately above, a scoring of six items within the CORE-OM has been developed, based on UK data, to give a QALY (Quality Adjusted Life Years) state based on an algorithm using those six items (Mavranezouli et al., 2011a, CORE-6D; 2011b, 2012). We will report distributions and psychometrics for the CORE-6D in Ecuador. To the best of our knowledge this will actually be the first replication/extension study of the CORE-6D scoring from the UK or elsewhere.
  - WGO measure.
    This is a hybrid qualitative and quantitative measure previously only been used in a study of adolescent school students in the UK (personal communication) hence this be only the second use of this measure. If person power can be found the events described by participants will be analysed using simple thematic analysis and associations of major themes with group, other demographics and dependent variable scores will be reported.

- P (Plan): *already spelled out in registration*
- D (Data): *ditto*

- A (Analyses): These will be divided into two tranches (see also Communication, below): analyses of baseline data and analyses of change when all data collection has been completed.
  All analyses will be conducted using R (R Core Team, 2020) and platform, R version and all non-base packages used will be reported either in publications or in publicly accessible supplementary material.
    - **Baseline analyses**
        - Demographic characteristics and baseline WGO and dependent variables scores will be reported. Primary presentations will be graphical with plots describing the diversity in the data (violin and scatter plots). Where available referential Ecuadorean data exist, sample statistics will be compared with referential data with CIs indicating precision of estimation. CIs for binary proportions will use Wilson's method and bootstrap CIs will be used for other variables, 95% intervals for CIs will be used throughout. Findings will be discussed with reference to likely systematic non-participation bias revealed in the comparisons.
          As well as the comparisons with other Ecuadorean data reported in the paper, CIs will allow readers to compare findings with data from other samples or populations of interest/concern to them taking into account likely imprecision of estimation in the study.
        - Where associations between baseline variables, including group, are statistically significant at $p < .01$ the interactions between these predictors and the dependent variables will be reported to throw light on the likely systematic associations between variables and the potential to confound one primary focus of interest: the differences between the two groups. The alpha value of .01 has been chosen *a priori* to select moderately strong associations and provide some control of the inevitably elevated rate of "false positives" above .05 that arises from multiple tests if an alpha of .05 is used.
            - With hindsight and having conducted analyses along these lines we are aware that there are many other ways of exploring these complexities in survey data and would particularly welcome offers from data analysts to explore other methods of exploration of these issues in this baseline dataset.
        - Baseline non-participation will be reported to the extent that the opportunistic recruitment process allows though this can only be in terms of potential number of students who might have opted into the study and then in terms of the numbers who start the online data collection but opt out before giving usable data.
        - For the CORE-OM and EQ-5D-3L internal reliability and factor analysis will be conducted (but not expecting neat factor structures). In addition, convergent validity correlations between the two measures, and with the EQ-VAS, will be reported, as usual with 95% CIs.
        - Distributions of CORE-OM, EQ-5D-3L and EQ-VAS scores and of WGO scores will be reported with CIs for centiles.
        - Baseline associations between those three dependent variables and the WGO scores will be reported with scattergrams and smoothed LOESS

regression lines and CIs.

- **Change analyses**
  - Patterns of post-baseline missingness will be reported. These will almost certainly show a preponderance of simple attrition: i.e. increasing non-response in waves 2 to 7, however, other patterns are also likely to be observed. As there are 64 possible further completion patterns (2^6) the missingness will be analysed in terms of simple dichotomy of "all completed" versus "at least one missing" and in terms of the total number of completions (1 to 7). These will be analysed against baseline demographic and dependent variable scores for evidence of any of these variables showing systematic associations with missingness. The associations will be reported testing against an alpha of .01 as in the baseline analyses but also with CIs for the associations.
  - One key focus is on how much scores on the three dependent measures, CORE-OM, EQ-5D-3L and EQ-VAS vary across the seven time points and particularly the question of whether that variation over time appears to vary systematically across individuals, i.e. that the within individual variance is not compatible with a null model of a shared population parameter showing only random individual differences? As is common in such studies (e.g., Cooper & McConville, 1990; McConville & Cooper, 1996; Murray et al., 2002) the within-participant standard deviation of the scores, "wpSD" across the seven assessment points will be reported and its distribution (histogram with individual points) plotted. The Levene test, an extension of the Bartlett test, will be used to test for heteroscedasticity, i.e. that null hypothesis of that this variance within group (i.e. here participants) is likely to have arisen by sampling vagaries from a shared population value. The Levene is said to give results in which the null hypothesis is rejected at the conventional alpha level of 0.05 more robustly for non-Gaussian distributions than does the Bartlett test.

    The wpSD is useful as it is a sensible statistic aggregating change across all occasions. However, in terms of possible implications for liability to gain from help, the aggregation across all measure completions risk hiding particularly high peak scores so two other within-participant change summary indices will be computed:
    - Signed maximum change from baseline score ("wpMaxChange").
    - Absolute maximum difference between any two completed scores ("wpScoreRange").

      These don't allow a formal test for homogeneity of change across participants but they do throw light on possible "spike" and "trough" changes.

      All three within-participant summary indices will be positively correlated with each other and this correlation will be reported and the findings carefully reported as correlated but addressing issues whose differences are of interest.
  - As noted, the baseline psychometric analyses, while useful guides to the internal reliability of the multi-item measures for comparisons across individuals at a single measurement occasion. However, the cross-sectional,

conventional test theory underlying that use of the internal reliability does not necessarily fit measurement of change within-participants. This will be explored as follows.

- The most common and stringent test of reliability of measurement of change within Classical Test Theory (CTT) of between person measurement is formal testing of longitudinal factorial measurement invariance which can test a sequence of levels of invariance configural, where the factor structure remains an acceptable fit to the model without fixing many population parameters to be the same on each occasion through to structural invariance in which factor intercorrelations/covariances, item loadings, item error variances and factor (latent variable) means can all be constrained to be equal across occasions without statistically significant misfit.
  We believe it is unlikely that either the CORE-OM scores, or the EQ-5D-3L, will fit a clean factor structure even at baseline, however it is possible, even with a poor baseline fit, to explore where and how badly longitudinal measurement invariance fails.
- Test-retest reliability is a paradigmatic and simple application of CTT to provide a guide to measurement stability. If all instability is measurement unreliability then test-retest reliability should be similar for all test-retest intervals and similar to internal reliability. This is implausible for state measures like the CORE-OM and EQ-5D-3L as changes in the latent variable *are* expected which makes observed changes a sum of those real changes, and error variance. However, stability indexed using the intra-class correlation coefficient across occasions can be used to compare test-retest with internal reliability. Where multiple test-retest intervals exist as here (21 test-retest intervals for those completing measures on all seven occasions) test-retest reliability can also be checked for stability, i.e. independence of interval, or for the more likely model in which test-retest reliability will decrease with increasing intervals between measure completions. The population model in which that reliability is zero can be tested for any test-retest interval as can the hypothesis that the test-retest stability is nearer to the internal reliability than it is to zero.
- However, as noted above, the CTT assumption that the within-participant change measurement model is the same as the between-individual model is highly implausible for mental health change measures. If that is accepted and that constraint abandoned these above checks on change measurement quality from within CTT cease to be definitive and could miss situations in which measures are useful measures of within-participant change, change which might be different across the items of the measure between individuals and having no shared factor structure between individuals. In that situation, an extreme deviation from the CTT model, change recorded across the items within each individual might still be usable reliable and valid indicators of within-

participant psychological change.
Abandoning the CTT model in this way still allows two simple explorations which can help understand more about the quality of measures as measures of within-participant change. These are currently new methods and seven occasions provide only minimal power for the second method but these will used as part of a wider project of inspection of the psychometrics of individual change.

- A measure can be useful as an aggregate measure of individual change if within-participant change is correlated across multiple items *but the correlated items being different for different individuals but not unique to individuals*. If this is the case, then Cronbach's alpha from item score changes (as opposed to scores at a point in time) is an indicator that, despite differences in the structure of change by individual, within-participant change is being indexed across items. This can be applied for any of the up to 21 score pairs.
- That still looks for some shared change across individuals. Dropping that measurement expectation entirely, but transferring CTT to within-participant change allows computation of the alpha for change within-participant. The parametric 95% confidence interval for an observed alpha of .6 from n = 7 (i.e. all seven occasions) for 28 items (the CORE-OM-NR) is from 0.0 to .92, i.e. reaching a point at which a model of no reliability of within-participant change is looking implausible. We will report within-participant Cronbach alpha values for all participants with full item data from all seven occasions for the CORE-OM-NR score and also, separately, for those with fewer occasions. Where the parametric lower confidence interval exceeds zero the first principal components of the within-participant PCA will be inspected and it is expected that these will be markedly different for different participants with these high personal change Cronbach alpha values, illustrating that within-participant change on a measure like the CORE-OM can be reliable be markedly different in structure between individuals. This can also be done for all 34 items making up the full CORE-OM. For the shorter scales/scores such as the EQ-5D-3L or the domain scores of the CORE-OM the number of items are not sufficient to give any reasonable power to find internal reliability with only seven measurement occasions and this model will not be explored for the shorter scales/scores.

- Does baseline group, student vs. non-student, appear to relate to change on the dependent variables? All change scores and distributions will be plotted allowing both individual values and central location statistics with CIs to be seen.
  All analyses will be firstly "intention to treat" (ITT), i.e. by baseline group.

However, many students will also hold jobs. How much this relates to scores on dependent variables will be explored by analysing with these "students-who-also-hold-jobs" as a group and by reclassifying them into the non-student group. These additional explorations will explore the sensitivity of any relationships to baseline group to these real world complexities.
It seems plausible that these "students-who-also-hold-jobs" are sufficiently different from either the students without any outside course employment or from the young people with jobs and no university study that this issue may need to be considered alongside the simple ITT analyses.
Some participants will probably change group during the year though these will probably be less numerous than "students-who-also-hold-jobs". Due to the timing of the data collection, students are probably more likely to terminate their courses during the year than are non-students to start a degree. Here again, sensitivity analyses will be conducted recomputing the analyses (a) omitting participants who change group at any point in the year and (b) omitting only those changing group whose change comes less than half way through the year and (c) reallocating those who change group to the group they fall in for more than half the year. However, the expectation is that the numbers changing group will be small and the impacts on the ITT findings small or negligible.
Recognising that these dependent variables have temporal variation and this variation is likely to be complexly patterned across the entire sample, the following analyses will be reported for the ITT baseline group comparison and the sensitivity analyses noted above.

- Group difference against mean scores across all non-missing scores from all seven occasions. This extends the analyses of baseline group differences and entirely ignores change over time.
- Group difference in linear trend over the time points using multi-level modelling (MLM). This allows a trend to be fitted both as a fixed effect whose interaction with predictors can be tested, i.e. assuming a single within-group population mean, and then also allowing a free effect, i.e. allowing that linear trend against time may vary between individuals. The MLM provides an individual (free) linear slope if scores at least two time points exist. However, sensitivity checks on effects of missingness will be added by restricting to sequentially larger numbers of time points.
- Group differences in within-participant SD of the scores (wpSD).
- Group differences in within-participant score difference from baseline (wpMaxChange).
- Group differences in within-participant score range (wpScoreRange).

- Do other demographic predictors relate to the dependent variables? Same methods as for group differences.
- To a substantial extent, the public health interest is more in high scores than in the full distribution or central locations. The distribution plots will have helped understand something of the upper tail of scores. However cutting points do exist for the dependent measures in Ecuador (i.e. classifications as “typically clinical" or "not typically clinical”: such as those derived in the Clinically Significant Change (CSC) paradigm (Evans et al., 1998; Jacobson &

Truax, 1991)). Hence, despite statistical power being lost by dichotomising, analyses will also report rates of change on the dichotomised scores.
As when analysing the continuous scores the relationship of change to ITT group, and to other baseline predictors will be reported. The public health issues are mostly about numbers starting and remaining above the CSC.

- Similarly, to help relate score changes to those observed with therapeutic interventions and to separate change with a probability of having occurred down to measurement unreliability alone (Reliable Change, RC: that above the Reliable Change Index, RCI) will be reported as the public health issues this addresses are about numbers showing reliable deterioration and and/or sustained reliable deterioration and about how these rates compare with such rates from samples receiving support or formal interventions.
- As well as all these analyses of scores and of score changes on the CORE-OM, EQ-5D-3L and EQ-VAS, this study seeks to discover information about levels of, and changes in, stressors/life events (LE) across the seven time points breaking exploration of this down across the following analyses.
- Is there evidence of systematic relationship between stress and dependent variable scores within-participant over time? This will be explored graphically and tested using an extension of the cross-lagged panel model (CLPM) introduced in Hamaker et al. (2015). That work introduced the full Random Intercepts CLPM (RI-CLPM) in which stable differences (random intercepts in the terminology of MLM) between participants across time in both predictor and dependent variables are partialled out before estimation of the autocorrelation and of the crucial cross-lagged regression coefficients.
  A modified model will be used in random intercepts are removed from the dependent variables in the model but the raw RGO scores are used. This model tests for effects of the WGO, i.e. life event, scores on the change in the dependent variable.
  Given that the WGO is "participant generated" i.e. invites the participant to name things that have happened in the previous week it, like the EQ-VAS, it is not amenable to psychometric explorations possible for nomothetic multi-item measures. That makes this exploration of a plausible predictive validity aspect of the WGO the only psychometric test of the measure.

- C: Communication
  General results will be sent to participants after the termination of data collection.
  The *a priori* analyses described above, and any *post hoc* analyses of unexpected findings, will be published in peer-reviewed journals and presented at relevant meetings with any *post hoc* analyses clearly marked and discussed as such.
  As noted above, reports will be of the baseline data and group differences and then, when the full data collection period has finished, of the change data.
  Brief reports of these publications will be disseminated in social media for knowledge of the community and private or public organizations interested in emerging adults' population.

References

Abelson, R. P. (1995). *Statistics as principled argument*. L. Erlbaum Associates.

Altman, D. G., & Moher, D. (2013). Declaration of transparency for each research article. *BMJ*, *347*(aug07 3), f4796–f4796. https://doi.org/10.1136/bmj.f4796

Åm, H. (2018). Ethics as ritual: Smoothing over moments of dislocation in biomedicine. *Sociology of Health & Illness*. https://doi.org/10.1111/1467-9566.12818

Cooper, C., & McConville, C. (1990). Interpreting mood scores: Clinical implications of individual differences in mood variability. *British Journal of Medical Psychology*, *63*(3), 215–225. https://doi.org/10.1111/j.2044-8341.1990.tb01614.x

Erhard, F., Jukschat, N., & Sammet, K. (2021). Lost in Translation? Openness, Reflexivity and Pragmatism as Guiding Principles for Cross-Language Qualitative Research. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, Vol. 22 No. 3 (2021): The Refiguration of Spaces and CrossCultural Comparison II. https://doi.org/10.17169/FQS-22.3.3722

Evans, C. (2012). The CORE-OM (Clinical Outcomes in Routine Evaluation) and its derivatives. *Integrating Science and Practice*, *2*(2). http://www.ordrepsy.qc.ca/pdf/2012_11_01_Integrating_SandP_Dossier_02_Evans_En.pdf

Evans, C., Connell, J., Barkham, M., Margison, F., McGrath, G., Mellor-Clark, J., & Audin, K. (2002). Towards a standardised brief outcome measure: Psychometric properties and utility of the CORE-OM. *British Journal of Psychiatry*, *180*(1), 51–60. Scopus. https://doi.org/10.1192/bjp.180.1.51

Evans, C., Margison, F., & Barkham, M. (1998). The contribution of reliable and clinically significant change methods to evidence-based mental health. *Evidence Based Mental Health*, *1*, 70–72. https://doi.org/0.1136/ebmh.1.3.70

Evans, C., Mellor-Clark, J., Margison, F., Barkham, M., Audin, K., Connell, J., & McGrath, G. (2000). CORE: Clinical Outcomes in Routine Evaluation. *Journal of Mental Health*, *9*(3), 247–255. https://doi.org/10.1080/jmh.9.3.247.255

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*(1), 102–116. https://doi.org/10.1037/a0038889

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*(1), 12–19.

Mavranezouli, I., Brazier, J. E., Rowen, Donna, D., & Barkham, M. (2012). Estimating a Preference-Based Index from the Clinical Outcomes in Routine Evaluation-Outcome Measure (CORE-OM): Valuation of CORE-6D. *Medical Decision Making*, *20*, 321–333. https://doi.org/10.1177/0272989X12464431

Mavranezouli, I., Brazier, J. E., Young, T. A., & Barkham, M. (2011a). Using rasch analysis to form plausible health states amenable to valuation: The development of CORE-6D from a measure of common mental health problems (CORE-OM). *Quality of Life Research*, *20*(3), 321–333. Scopus. https://doi.org/10.1007/s11136-010-9768-4

Mavranezouli, I., Brazier, J. E., Young, T. A., & Barkham, M. (2011b). Using Rasch analysis to form plausible health states amenable to valuation: The development of CORE-6D from a measure of common mental health problems (CORE-OM). *Quality of Life Research*, *20*(3), 321–333. https://doi.org/10.1007/s11136-010-9768-4

McConville, C., & Cooper, C. (1996). Mood variability and the intensity of depressive states. *Current Psychology*, *14*(4), 329–338. https://doi.org/10.1007/BF02686921

Molenaar, P. C. M. (2004). A Manifesto on Psychology as Idiographic Science: Bringing the Person Back Into Scientific Psychology, This Time Forever. *Measurement: Interdisciplinary Research & Perspective*, *2*(4), 201–218. https://doi.org/10.1207/s15366359mea0204_1

Murray, G., Allen, N. B., & Trinder, J. (2002). Longitudinal investigation of mood variability and the ffm: Neuroticism predicts variability in extended states of positive and negative affect. *Personality and Individual Differences*, *33*(8), 1217–1228. https://doi.org/10.1016/S0191-8869(01)00217-3

Paz, C., Adana-Díaz, L., & Evans, C. (2020). Clients with different problems are different and questionnaires are not blood tests: A template analysis of psychiatric and psychotherapy

clients' experiences of the CORE-OM. *Counselling and Psychotherapy Research*, *20*(2), 274–283. https://doi.org/10.1002/capr.12290

Paz, C., Mascialino, G., & Evans, C. (2020). Exploration of the psychometric properties of the Clinical Outcomes in Routine Evaluation-Outcome Measure in Ecuador. *BMC Psychology*, *8*(1), 94–105. https://doi.org/10.1186/s40359-020-00443-z

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.r-project.org/

Spiegelhalter, D. (2019). *Art of Statistics: Learning from Data.* Penguin Books, Limited.

Wagenmakers, E.-J., Sarafoglou, A., Aarts, S., Albers, C., Algermissen, J., Bahník, Š., van Dongen, N., Hoekstra, R., Moreau, D., van Ravenzwaaij, D., Sluga, A., Stanke, F., Tendeiro, J., & Aczel, B. (2021). Seven steps toward more transparency in statistical practice. *Nature Human Behaviour*, *5*(11), 1473–1480. https://doi.org/10.1038/s41562-021-01211-8