Summarizing the outcomes of a multiverse analysis

Kuipers, Charlotte

Summarizing the outcomes of a multiverse analysis

Charlotte Kuipers

Master's Thesis Psychology,
Methodology and Statistics Unit, Institute of Psychology
Faculty of Social and Behavioral Sciences, Leiden University
Date: 07-12-2021
Student number: 2346982
Supervisor: Dr. T.D.P. Heyman (internal)

Abstract

The multiverse analysis can be used as a way of assessing the influence of different analysis choices that could reasonably be made by researchers, instead of only presenting the result of one research 'path' as is often done in studies. While the multiverse analysis increases transparency about the results, it is still unclear how researchers can best summarize the results of this analysis more formally. Moreover, as far as we are aware, no previous studies have examined how the multiverse analysis performs under different research conditions. In this study, we simulated data under different research conditions. In addition, we built a generic multiverse analysis that was used to analyze this data. Two methods were used to summarize the results of this analysis, namely the mean p-value and the harmonic mean p-value (HMP). The results of this study showed that the mean p-value may be the preferred summarization method, as it provides a more conservative estimate of the different paths in the multiverse and has less false-positive results than the HMP in a situation where data was simulated under the null hypothesis. In addition, our study shows that the summarization methods of our multiverse analysis are robust against variations regarding the number of variables that are part of the analysis, the amount of missing data in a dataset and changes in the correlation between variables. However, the summarization methods in our multiverse were not robust against underpowered data. Only if the different research paths in our multiverse analysis had adequate power, the HMP was generally able to find a significant result in at least 90% of cases. However, future research is needed to see if these results can be replicated, since the definition of a generic multiverse analysis may differ depending on the research field.

Table of content

Introduction

1.1. About the multiverse analysis

During the researching process, there are several decisions that researchers can make regarding how to handle their data. For example, researchers may opt to combine or remove certain variables, or to handle missing data and/or outliers in a certain manner (e.g., using listwise deletion and removing outliers). These seemingly arbitrary choices are referred to as *researcher degrees of freedom* by Simmons, Nelson and Simonsohn (2011). The authors also describe how selectively using these researcher degrees of freedom could increase the rate of false-positive results. It should be noted that this increased false-positive rate would only occur if researchers would selectively use these researcher degrees of freedom and would not be transparent about their research process. For example, the false positive rate would only increase if researchers decided to run different analyses with different outlier criteria, but only report the result for one of these criteria and make it seem as if this was the only analysis that was performed.

Some suggestions have been made to limit the impact of selectively using these researcher degrees of freedom. For example, Simmons et al. (2011) suggest (amongst other things) that authors should report the results of all of their experimental conditions, including the failed ones, in order to avoid only reporting the results that 'worked' for their hypothesis. In addition, they suggest that if observations are removed from the analyses, the researchers must also report the results including these observations. These suggestions would limit the ability of authors to selectively use these researcher degrees of freedom and it would also increase transparency about the analyses and obtained results. However, an even better approach of dealing with these researcher degrees of freedom would be the use of a *multiverse analysis*. Essentially, one could argue that the multiverse analysis builds on the suggestions made by Simmons et al., as it is a principled way of assessing the influence of these analyses choices (e.g., removing versus not removing outliers from the analysis).

In 'traditional' research, usually only one result is reported that is (potentially) the product of selectively using these researcher degrees of freedom. Meanwhile, a multiverse analysis can be used to investigate the outcomes of a *multiverse* of possible datasets, thus providing a more transparent result. Instead of performing one set of analyses based on one constructed dataset that is (potentially) the product of researcher degrees of freedom, the multiverse analysis is able to show what would happen if different choices regarding the handling of variables and/or analyses would have been made (Steegen, Tuerlinckx, Gelman,

& Vanpaemel, 2016). For example, Steegen and colleagues performed a multiverse analysis using an existing dataset from two studies that investigated the influence of fertility on religiosity and political attitudes. The data was collected by means of a questionnaire and involved the variable 'relationship status'. The researchers would then classify the participants either as single or as being in a committed relationship. However, one could argue that participants who selected 'dating' as their relationship status could reasonably be classified as being in either group. Thus, instead of arbitrarily deciding on one research path (e.g., classifying all dating participants as being single) and reporting a single result as is mostly done in traditional research, Steegen and colleagues decided to perform a multiverse analysis which incorporated all of the choices that could reasonably have been made in regards to this variable. In total, three paths/options were created for this variable. In one of the paths, participants who were dating were classified as being single, while in the second path they were classified as being in a committed relationship. The third path discarded participants who were dating from the analysis entirely, due to the ambiguity of this relationship status. It should be noted that a multiverse analysis can also incorporate paths that involve analysis choices. For instance, this could happen when a researcher decides to use different methods of handling missing data. Here, instead of reporting a single result as is done in traditional research (e.g., using listwise deletion), the researcher would also present the results for other methods that handle missing data (e.g., pairwise deletion or forms of multiple imputation).

1.2. Aim of the current study

While the multiverse analysis can show the results of the different research paths, it is still unclear how researchers can best summarize and interpret the results. Researchers can present/visualize the results of the multiverse analysis through creating histograms or a heatmap which show the resulting p-values of each research path, similar to what Steegen et al. (2016) did in their study. However, interpreting the results is mostly done by eyeballing the p-values in each of the paths, which might prove to become more difficult as the multiverse becomes larger. In addition, as far as we are aware, no previous studies have looked into the performance of the multiverse analysis as a whole.

In the current study, we will address both of these issues. We built a generic multiverse (i.e., a 'basic' multiverse analysis which can be applied in different research fields) in which one-sample t-test were performed. The data for this generic multiverse was also simulated, so that we knew if there truly is an effect that needed to be found by the

multiverse analysis or not. To measure the performance of this generic multiverse under different research situations (i.e., whether the data was simulated to have an effect or to have no effect), we followed a suggestion that was offered by Steegen et al. (2016) about how to summarize the results of a multiverse analysis more formally. Steegen and colleagues suggested using the mean p-value that resulted from the multiverse. This meant that the arithmetic mean was calculated from all the p-values in the different paths of the multiverse, thus resulting in one single mean p-value for the entire multiverse that was performed on a dataset. Steegen et al. argue that this mean p-value can be considered as the p-value that results from a hypothetical pre-registered study in which the conditions were chosen at random from all of the possibilities in the multiverse.

In addition to using the mean p-value, we used the harmonic mean p-value (HMP) as a way of summarizing the results of the multiverse. The latter can be used in situations where the results are dependent upon each other, as is the case with the multiverse analysis (Wilson, 2019). Wilson suggests the HMP as a better way of controlling for the familywise error rate, since the Bonferroni correction might be too conservative in situations where the results are dependent upon each other. When the HMP is significant, it means that at least one of the p-values in the multiverse was significant. If the HMP provides a non-significant result, it means that none of the research paths yielded a significant result.

We used these two metrics as a way of measuring the 'performance' of the multiverse analysis under different research conditions. In the current study, performance was defined as the proportion of mean p-values and HMP's that reflected the true situation. Thus, in a situation where we would know that there is a true effect, we would expect the mean p-values and HMP's to be significant most of the time. In addition, a research condition in which there was no effect to be found (i.e., a situation where the null hypothesis reflects the true situation) was investigated in the current study as well. With this approach, we were able to investigate how the two ways of summarizing the multiverse perform compared to each other, while we would also be able to investigate the performance of the multiverse under different research conditions.

Method

2.1. Building the generic multiverse

One can construct a multiverse analysis in many different ways. For example, some researchers opt to create the multiverse by basing the different pathways on arbitrary choices that could have been made by the researcher (i.e., removing or combining certain variables) (Cesario, Johnson, & Terrill, 2018; Dejonckheere et al., 2018), while other researchers create the paths for the multiverse according to recommendations and common practice in their own field (e.g., using a different scaling method that is often used in their field of research) (Denny & Spirling, 2018; Liebst et al., 2019). It was not feasible to account for all this variation in the current study. Instead, we created a generic multiverse which incorporated options that we believed to be common in most research (within psychology).

Two functions were key in creating the multiverse for this study, both of which were created in R. These functions are called create_data and multiverse_ttest. The first function, which is called create_data (see Appendix B), uses the simsem package to generate data from a confirmatory factor analysis (CFA) model that is specified by the user (Pornprasertmanit, Miller, Schoemann, & Jorgensen, 2020). An example of this CFA-model can be found in Figure 1.

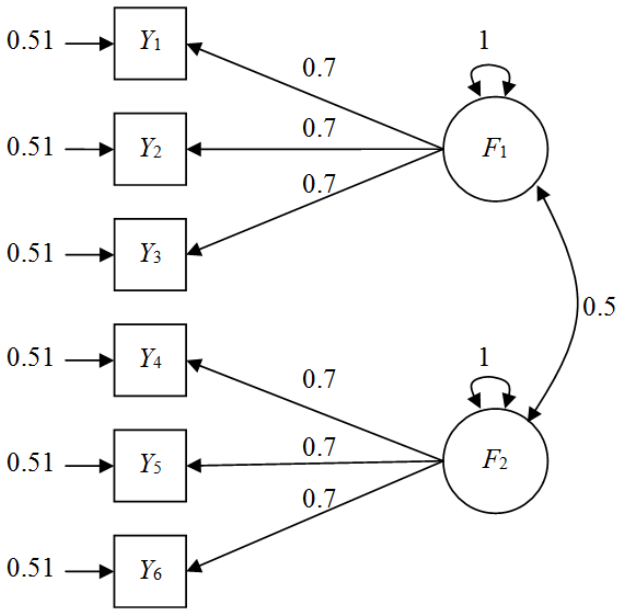

Figure 1. Example of CFA model that can be used to generate data. Retrieved from simsem Github wiki (Simsem, 2021).

The create_data function offers the user options to specify the sample size (i.e., the number of observations), the number of factors, the number of indicator variables on each factor, the mean of the indicator variables, whether the data should have missing data, and the proportion of missingness. It should be noted that while there is an argument to specify the number of factors in the function, this was not entirely operational at the time of writing. Instead, the data will always be drawn from a model with two factors, but only the first factor and corresponding indicator variables were used for the multiverse in the current study. In addition, note that there is an additional argument in the function to specify the correlation between the factors, but this was not used in the current study since we focused solely on the first factor and corresponding variables with the multiverse_ttest function. However, a separate function called multiverse_cor was made which is able to examine correlations between the variables. While it was outside the scope of the current study, this function is shared in Appendix C so that other researchers may use this function if they wish to.

A hypothetical research question for the first factor and corresponding indicator variables in our multiverse analysis could be: 'Do people in the province of South Holland differ in intelligence from the population (i.e., the rest of the Netherlands)?'. In this example, the latent factor would be intelligence and the variables could each be different tests that attempt to measure intelligence and which all have the same population mean. After all, deciding which test to use could be considered an arbitrary choice/ researcher degree of freedom and could thus be part of our multiverse analysis. One could also think about the first factor and corresponding indicator variables from a different perspective with regards to our multiverse analysis. For example, a researcher could attempt to measure general knowledge (latent factor) by means of a single open-ended trivia questionnaire. Here, the indicator variables could be different ways of measuring correct/incorrect answers to the trivia questionnaire as this too, could be considered an arbitrary choice.

The second key function for this study is the multiverse_ttest function (see Appendix B), which is used to obtain the results of the multiverse analysis. In this function, the user is able to specify how they would like to manage missing data and outliers. Since our aim was to build a multiverse that was as generic as possible and could be used as a basis for multiverse analyses in different research fields, we looked at the literature to find the options that were most often used in practice. In multiverse_ttest there are a total of 4 options that handle missing data in a certain manner. The option 'none' can be used if there are no missing data generated. If there are indeed missing data in the dataset, users can opt to use listwise deletion, pairwise deletion, and multiple imputation. The terms listwise and pairwise

might be a bit confusing in this situation with a one-sample t-test. Since a one-sample t-test only uses one variable for its analysis, listwise and pairwise would theoretically yield the same result as the missing data point(s) would just be ignored when running the analysis. However, this is different in the current study. With listwise deletion, the entire case (i.e., row) is removed if there is at least one missing data-point on any of the indicator variables of the first factor. Meanwhile, pairwise deletion does not delete the entire row. For example, if there is a case which has a missing data point at Y1, but not at any of the other variables, pairwise deletion opts to remove that case from analyses where Y1 is involved but still uses it for analyses involving the other indicator variables. In contrast, listwise deletion would remove not only the Y1 from the analysis, but also the values of the other variables of that case from further analysis. Multiple imputation is performed with the default settings of the mice function from the R package 'mice' (Van Buuren & Groothuis-Oudshoorn, 2011). Note that there are numerous other ways of handling missing data that could also have been chosen for the generic multiverse (Kang, 2013). However, there were some reasons as to why these three options were ultimately chosen. For instance, a review by Dong and Peng (2013) showed that listwise and pairwise deletion were the most used methods by researchers in cases of missing data. While this may be true, using listwise or pairwise deletion comes with some disadvantages (e.g., reduced power) which is why it has often been argued that multiple imputation is a better way of handling missing data (Dong & Peng; Schafer & Graham, 2002). Thus, in the generic multiverse we decided to incorporate listwise and pairwise deletion since these were most likely to be used in practice, while we also incorporated multiple imputation as a more ideal option of handling missingness that researchers may be aware of.

Furthermore, (potential) outliers can be handled in three different ways by the function. If the user selects 'none' nothing is done to the outliers and they remain in the dataset. In fact, with this path in the multiverse in addition to paths that would remove outliers, we would be following the advice previously mentioned by Simmons et al. (2011), since we would also report the result if no observations (in the form of outliers) are removed from the analyses. The other two options do focus on identifying and removing outliers however. Bakker and Wicherts (2014) reported how identifying outliers based on z-scores is by far the most common approach among researchers. They reported that the cut-off point where researchers deleted outliers ranged from approximately 1 to 10, but that the median was 3. Thus, we implemented this threshold in our multiverse_ttest function in the argument named 'z_remove'. With this option, the z-scores are calculated for each variable and if the z-

score is above 3 or below -3, the score in the original dataset is set to be missing. The final option of handling outliers is called 'IQR_remove'. If the user selects 'IQR_remove' the outliers are set to be missing if they are below 1.5 times the interquartile range of the first quartile or above 1.5 times the interquartile range of the third quartile. While this method of outlier detection was used in only 3.6% of the articles that Bakker and Wicherts reviewed in their study, they still recommend it as a better way of identifying outliers and mention that it is often taught to psychology students in their statistics textbooks.

In the current study, we focused on datasets that were generated to have missing data, as this occurs more often in research compared to not having any missing data (Dong & Peng, 2013). In addition, we simulated datasets that had either 3 or 5 indicator variables, as will be explained further in the next section. Thus, for the dataset with 3 indicator variables, the full multiverse would yield: 3 (indicator variables) × 3 (options for handling missing data: listwise, pairwise and multiple imputation) × 3 (options for handling outliers: none, z_remove and IQR_remove) = 27 p-values. While the multiverse with 5 indicator variables would yield: 5 (indicator variables) × 3 (options for handling missing data: listwise, pairwise and multiple imputation) × 3 (options for handling outliers: none, z_remove and IQR_remove) = 45 p-values.

2.2. Manipulating the research conditions

In this simulation study, we were able to create research conditions that one could find in reality. For example, by setting the indicator means to 0 in create_data we were able to create a situation where the data was drawn from a model where the null hypothesis is true, since we tested against zero ($\mu = 0$). In addition, we used Cohen's d effect sizes to create situations where the simulated data would be consistent with the alternative hypothesis. This was done by setting the indicator means to 0.2, 0.5 or 0.8 respectively, according to the rules of thumb for small, medium and large effect sizes by Cohen (1988). Note that setting the indicator means to these values in order to reflect Cohen's d effect sizes was only possible because we tested against $\mu = 0$ and simulated the data through a normal distribution in which the standard deviation was set at one.

This left us with a total of four research situations which were investigated in the current study. In addition, more manipulations to these research conditions were made so that they reflect real research conditions. Note that Table 1 provides an overview of the different research situations and their manipulations.

Table 1

An overview of the four research situations and their manipulations

	Sample size	Missingness proportion	Number of indicator variables	Loadings
Situation 1: Cohen's d = 0.0	198 and 40	5% and 15%	3 and 5	$\sqrt{0.5}$ and $\sqrt{0.3}$
Situation 2: Cohen's d = 0.2	198 and 40	5% and 15%	3 and 5	$\sqrt{0.5}$ and $\sqrt{0.3}$
Situation 3: Cohen's d = 0.5	33 and 40	5% and 15%	3 and 5	$\sqrt{0.5}$ and $\sqrt{0.3}$
Situation 4: Cohen's d = 0.8	14 and 40	5% and 15%	3 and 5	$\sqrt{0.5}$ and $\sqrt{0.3}$

The first manipulation involved sample size, and two options were created. The first option was based on a power calculation, where we calculated the sample size that was needed to find the previously described small, medium and large Cohen's d effect sizes of Situation 2, 3 and 4 in a single pathway. The power was set at 0.8, which resulted in a sample size of 198 for Situation 2, a sample size of 33 for Situation 3, and a sample size of 14 for Situation 4. Note that the power calculation took place without taking the missing data and possible removal of outliers into account. As a result, the effective power is a bit lower, since our multiverse has paths that use listwise deletion, thus deleting some cases. The same is true (although to a lesser extent) for the paths that used pairwise deletion. For example, in Situation 2 we created data from a model where we know that the alternative hypothesis of 0.2 is 'true'. This led to a sample size calculation of 198, but because we also simulated missing data, some data was removed due to the paths in the multiverse that involved listwise and pairwise deletion, thus leading to slightly underpowered analyses in these pathways. In addition, we chose to use a sample size of 40 as a second option, since this appeared to be the median sample size in psychological research, thus creating another realistic option that might be found in other research (Marszalek, Barber, Kohlhart, & Holmes, 2011). With a sample size of 40, the power would set at approximately 0.235 in Situation 2, 0.869 in Situation 3 and 0.999 in Situation 4 (all values were rounded of at the third decimal). It should be noted that Situation 2 (Cohen's d of 0.2), Situation 3 (Cohen's d of 0.5) and Situation 4 (Cohen's d of 0.8) all incorporated these two options regarding sample size, but that this was different for Situation 1 (null hypothesis is true). Naturally, power could not be calculated in this situation since there is no effect to be found. Furthermore, the sample size should not matter in situations where the null hypothesis is true. Thus, the decision was made to use 198 as a sample size, which is the same as the power calculated sample size in

Situation 2. In addition, 40 was used as a sample size as well to keep the same format as the other research situations which make the results easier to compare.

We also created two options for the amount of missing data that is present in the dataset (assuming missing completely at random (MCAR), see Appendix D for a justification). One option was used to generate 5% missingness, and the other option was used to generate 15% missingness. Research has shown that having 5% or less missing data within a dataset is often inconsequential and does not lead to biased estimates (Schafer, 1999). It should be noted that with MCAR, ignoring the missingness (i.e., use listwise deletion) would not introduce bias in the results since the missingness is completely at random (Dong & Peng, 2013). However, it would still reduce the power of the statistical analyses since there are less cases to work with. In the current study, we chose to include a missingness proportion of 5% to reflect what researchers could consider to be an unproblematic amount of missingness. In addition, we chose to incorporate an option with 15% missingness as well. This is because Enders (2003) stated in his simulation study that this proportion of missingness appears to be common in educational and psychological research. While Enders mentions that this is not the proportion of the whole dataset but rather the proportion of missingness on a subset of the variables in the dataset, the decision was made in the current study to simulate the missingness across the whole dataset (i.e., the indicator variables on the first factor). This is because of logical considerations. For example, it would be highly unlikely to find a real situation where there would only be missingness on two of the three variables in a multiverse analysis. This would mean that data entry mistakes or technical issues only occurred on two of the three IQ tests, or two of the three coding keys of the trivia questionnaire (in a situation where there are three indicator variables on the first factor). We could have made the decision to simulate the missingness proportion on all the variables. However, this would have severely limited our ability to measure the performance of the multiverse analysis accurately. As will be explained in further detail in the next section, one way we measured performance was by calculating the mean p-value out of the 27 or 45 p-values from the multiverse. However, the options in the multiverse that included listwise deletion (and to a lesser extent pairwise deletion) could potentially introduce a severe skewness on this mean p-value. For example, if we take a sample size of 100 and simulate three indicator variables with a missingness rate of 15%, there is a possibility that 45 cases could be removed from the analysis with the paths that include listwise deletion. Thus, we would be losing power. In conclusion, we believed that simulating missing data across the whole dataset instead of per variable would be the better choice here.

Another manipulation involved the number of indicator variables on the first factor. We decided to make two options, one where there are three indicator variables and one where there are five. We believed that three indicator variables was a feasible amount for creating a multiverse, and that most researchers could create a multiverse analysis with this number of variables in their study. Using five indicator variables in a study might not always be a justifiable choice however. For example, researchers could still use five different ways of measuring correct/incorrect answers to a trivia questionnaire, with each of the different ways being a justifiable choice. Meanwhile, using five different IQ tests might be a bit much, but using three might still be feasible. It should be mentioned that using several indicator variables to measure a construct is not necessarily the goal when one wishes to create a multiverse analysis. In fact, using several indicator variables/paths while there clearly is only one justifiable choice in measuring the construct would be ill-advised. In practice however, it is likely that there are a number of indicator variables that researchers could use in their multiverse analysis. In addition, note that a large number of indicator variables (i.e., 20) might be possible in some specific research situations. For instance, this could occur if a researcher wanted to investigate whether certain attributes of a neighborhood (measuring neighborhood quality) differ from the rest of the city. This is a broad research question which could involve a lot of different variables. For example, the researcher could look at crime rate, access to public transport, school drop-out rates, presence of recreational spaces and many more variables in their multiverse analysis. The researcher could also decide to code these variables differently (e.g., divide crime data in a variable measuring non-violent crime and a variable measuring violent crime), adding to the number of variables that are in the multiverse even more. However, the aim of the current study is to make a generic multiverse. Therefore, using three and five variables seemed to be reasonable, while also showing some contrast through a smaller multiverse option of three variables (resulting in 27 pathways) and a bigger multiverse option of five variables (resulting in 45 pathways).

Finally, the loadings of the indicator variables were changed to correspond with high and medium correlations (i.e., 0.5 and 0.3) between the variables according to Cohen's rules of thumb (Cohen, 1988). A third option that reflects a small correlation was left out, since one could argue that if an indicator variable correlates relatively poorly with (one of the) other variables, it should not be a part of the multiverse at all. The loadings for the indicator variables were created by taking the square root of the medium or large correlations according to Cohen's rules of thumb.

To summarize, in the current study we used four research situations that were manipulated further. Situation 1 (the null hypothesis is true), Situation 2 (Cohen's d of 0.2), Situation 3 (Cohen's d of 0.5) and Situation 4 (Cohen's d of 0.8), each lead to 2 (options for sample size) × 2 (options for missingness proportion) × 2 (options for the loadings of the indicator variables) × 2 (options for the number of indicator variables on the factor) = 16 variations.

2.3. Measuring performance

To measure the performance of the multiverse, we used the previously described mean p-value and HMP. For the mean p-value, this meant that instead of reporting all 27 or 45 p-values for each of the research paths, the mean was calculated which resulted in a single p-value for the entire multiverse that was performed on a dataset. The other form of summarizing the multiverse analysis was the HMP. If this value was significant, it meant that at least one of the 27 or 45 paths in the multiverse were significant. Meanwhile a non-significant HMP meant that none of the paths in our generic multiverse yielded a significant result.

As was briefly described in the introduction, we then measured the performance/statistical power of the generic multiverse by calculating the proportion of results that reflected the true situation from the total amount of simulated datasets. For example, in Situation 2 we simulated data from a model where we would know that in reality, the mean of the indicator variables is 0.2. By calculating the proportion of mean p-values and HMP's that showed significant results (and thus reflecting the true situation), the performance/power of the multiverse could be calculated. An alpha level of 0.05 was used to indicate significant results. Of course in Situation 1 where the null hypothesis is true, we would expect mostly non-significant results instead of significant ones. One could not accurately speak of statistical power here either, since there is no effect to be found under the null hypothesis. Instead, the term 'true negatives' will be used when speaking about the proportion that is accurately classified as showing no significant results in Situation 1.

We simulated each research situation and corresponding manipulations 1000 times which resulted in 16000 mean p-values and HMP's for Situation 1, 2, 3 and 4 each. However, it should be noted that with some manipulations, no missing data was generated even though this was specified in the create_data function (by either 5% or 15% missingness). This would occur in situations with relatively small sample sizes, namely the sample sizes of 40, 33 and 14. To acquire our data using the create_data function, a dataset without missingness was first

generated from a CFA-model. Then, in a next step, the missing data was generated with the ampute function from the mice package, which resulted in some variation regarding the exact number of missing data points that were generated (Van Buuren & Groothuis-Oudshoorn, 2011). For example, with a sample size of 40 and a 5% missingness proportion, we would expect the create_data function to simulate approximately 2 missing data points across the indicator variables on the first factor. However, due to some slight variation around this number created by the ampute function, we could end up with a dataset in which no missing data points were generated. When this occurred, we ran the multiverse_ttest function with 'none' specified as a way of handling missing data instead of the 'listwise', 'pairwise' and 'mi' arguments that we would normally use.

All analyses were performed in R version 4.0.2. with the packages 'harmonicmeanp' (Wilson, 2019), 'knitr' (Xie, 2020), 'mice' (Van Buuren & Groothuis-Oudshoorn, 2011), 'psych' (Revelle, 2020), 'pwr' (Champely, 2020) and 'simsem' (Pornprasertmanit et al., 2020) installed. The corresponding R code that was written in R Markdown can be found in the Appendix B.

Results

Two tables were made to present the results of each research situation in order to increase transparency. For each research situation, the first table is used to present the results solely of the datasets in which missing data was actually generated. Meanwhile, the second table is used to present the results of all 1000 datasets. Note that a column reporting the number of observations was added to each table. This way, we are able to observe how often it occurred that a dataset was generated with or without missing data. For example, if the number of observations is 1000 for a manipulation in the first table (describing solely the datasets that were generated with missing data), we would know that all of the datasets for a certain research manipulation were generated to have missingness since we simulated each manipulation of a research situation 1000 times. If this number is lower, say 800 for example, we would know that in 200 iterations no missingness was generated in the datasets for that specific research manipulation. The second table will always have 1000 observations for each manipulation, since this table presents the combined results. Finally, note that in the first table in each research situation, the proportion is always calculated from the total number of datasets that were generated to have missing data for that specific combination of research manipulations. In the second table, the proportion is always calculated from the total number of generated datasets (i.e., 1000), regardless of whether they were generated with or without any missing data. In addition, note that the tables regarding the research results can be found in Appendix A.

3.1. Situation 1 (the null hypothesis reflects the true situation)

Table A1 provides an overview of the results for the datasets that were generated to have missingness. Meanwhile, Table A2 presents the combined results of the datasets that were generated both with and without missing data. From these tables, we can see that both the mean p-value and the HMP perform relatively well seeing how the proportion of finding a true negative result is always above 0.900. However, the mean p-value reflects this true negative result more often compared to the HMP in each of the different manipulations. The true negative rate ranges from 0.990 to 0.999 for the mean p-value (also in the combined datasets) and between 0.915 to 0.944 for the HMP (between 0.918 and 0.944 for the combined datasets). In a single pathway analysis, we would expect a true negative rate of 0.95 because 0.05 was used as our alpha level. When we compare our results to this true negative rate of a single pathway analysis, we see that the mean p-value reflects a true

negative result more often by comparison. Meanwhile, the HMP reflects this result less often, thus leading to more false-positive results.

In addition, Table A1 shows us that the datasets were always simulated to have missing data when a sample size of 198 was used, since the number of observations are 1000 in these combinations. In most cases, missing data was also simulated if a sample size of 40 was used. However, the exact number differs somewhat depending on whether a missingness proportion of 5% or 15% was used in combination with a sample size of 40, with the combination with 15% missingness ranging from 997 to 999 in the number of observations and the combinations with 5% missingness ranging from 852 to 895. This was to be expected, seeing how a missingness rate of 5% is more likely to result in a dataset without any missingness in small(er) sample sizes compared to a 15% missingness rate. Furthermore, when we again look at the results of both tables, we see that using either a sample size of 198 or 40 does not appear to have any effect on the rate at which the mean p-value detects a true negative result. However, the HMP does appear to have a somewhat lower proportion of showing a true negative result in combinations which used a sample size of 40 compared to combinations which used 198 as a sample size.

Overall, there does not seem to be any substantial impact on the results stemming from whether a 5% or 15% missingness rate was used or whether three or five indicator variables were used. Differences are also slim between the two loading options. However, in Table A2 a pattern can be found regarding the loadings for the mean p-value. While keeping the other manipulations such as sample size in the combinations at a constant, the mean p-value always has a higher proportion of finding a true negative result when the loadings are based on a medium correlation (i.e., $\sqrt{0.3}$) compared to when the loadings are based on a large correlation (i.e., $\sqrt{0.5}$). While there is a pattern, the differences can be described as rather small since the largest difference is 0.007. No such a pattern can be found for the HMP. Nor can any other patterns be found regarding the results in both tables.

3.2. Situation 2 (the alternative hypothesis of 0.2 reflects the true situation)

Contrary to Situation 1, the HMP now outperforms the mean p-value which can be seen in Table A3 (describing the datasets that were generated with missing data) and Table A4 (describing the combined results). The HMP performs better in terms of power, as it has a higher proportion of finding a significant result that reflects the true situation in each of the combinations compared to the mean p-value. The power of the mean p-value ranges between

0.069 and 0.736 (this range is the same for the combined datasets). For the HMP this range is between 0.298 and 0.939 (between 0.302 and 0.939 for the combined datasets). For both summarization methods, the range is quite large. It clearly makes a difference as to whether a sample size of 198 or 40 was used. If a sample size of 198 is used, the range of the power is between 0.689 and 0.736 for the mean p-value and between 0.890 and 0.939 for the HMP (these ranges are the same for the combined datasets). In the combinations with a sample size of 40, the power ranges between 0.069 and 0.125 for the mean p-value (between 0.069 and 0.123 in the combined datasets) and for the HMP this ranges between 0.298 and 0.365 (between 0.302 to 0.366 in the combined datasets). Note that in each combination in Table A3, the exact same number of datasets were generated to have missing data as in Situation 1. This explains why Table A3 and Table A4 have exactly the same results in the combinations which used a sample size of 198, since the number of observations were always 1000 in these combinations. In addition, in a single pathway analysis, the power was 0.8 if a sample size of 198 was used and 0.235 if a sample size of 40 was used. In our tables, we see that the mean p-value has less power (by approximately 0.1, for each sample size) than a single pathway analysis in each combination. Meanwhile, the HMP has more power in each of the combinations compared to a single pathway analysis, as the proportion of finding a significant result is higher by approximately 0.1.

Finally, the amount of indicator variables, whether a 5% or 15% missingness rate was used, or whether the loadings were based on a medium or large correlation did not seem to have much impact on the results. There does appear to be a pattern however, where the HMP is somewhat higher in power if loadings were used that were based on a medium correlation compared to a large correlation (while keeping everything else constant). In addition, a pattern can be found where in the combinations with a sample size of 40, the power of the mean p-value is slightly lower if a 15% missingness rate is used compared to 5%. For the HMP, this is the other way around, since the power becomes slightly higher if a sample size of 40 is combined with a 15% missingness rate. While there is a pattern, the differences are very limited, seeing how the largest difference is 0.019 for the mean p-value (0.018 for the combined datasets) and 0.025 for the HMP (0.020 for the combined datasets).

3.3. Situation 3 (the alternative hypothesis of 0.5 reflects the true situation)

In Table A5 (presenting the results solely of the datasets that were generated with missingness) and Table A6 (presenting the combined results), we can see that the HMP has consistently higher power than the mean p-value, with the mean p-value ranging from 0.690

to 0.854 (between 0.692 and 0.848 for the combined datasets) and the HMP from 0.880 to 0.980 (between 0.886 and 0.982 for the combined datasets). Overall, the power of both the HMP and the mean p-value seems to be somewhat higher in combinations where a sample size of 40 was used compared to combinations which had 33 as a sample size. If a sample size of 33 was used, the mean p-value ranges between 0.690 and 0.735 (between 0.692 and 0.747 for the combined datasets) and between 0.880 and 0.938 for the HMP (between 0.886 and 0.938 for the combined datasets). When a sample of 40 was used, the power of the mean p-value ranges from 0.803 to 0.854 (between 0.802 and 0.848 for the combined datasets) and the HMP ranges between 0.935 and 0.980 (between 0.935 and 0.982 for the combined datasets). In a single pathway analysis, the power would be 0.8 for the sample size of 33 and 0.869 for the sample size of 40. From the results, we can see that the power of the mean p-value is lower by approximately 0.1 in the combinations which use 33 as a sample size compared to the power of a single pathway analysis. In the combinations which used 40 as a sample size, the power of the mean p-value approximates the power of a single pathway analysis rather well, with the lowest power of the mean p-value being 0.803 (0.802 for the combined datasets), and the highest value being 0.854 (0.848 for the combined datasets) in these combinations. Meanwhile, the HMP exceeds the power of a single pathway analysis in each combination. The HMP scores higher by approximately 0.1 in each combination compared to a single pathway analysis with regards to the respective sample size.

Contrary to the previous research situations, Table A5 shows that datasets without any missingness were generated in each of the different combinations (i.e., not a single combination resulted in 1000 observations). Although Table A5 shows that more datasets with missing data were create in situations where a missingness proportion of 15% was used (ranging from 990 to 999 observations) compared to the situations where 5% was used (ranging from 811 to 895 observations). In addition, in both tables we again see a difference in results stemming from using either a 15% or 5% missingness rate. The power of the mean p-value is either equal or slightly lower in combinations where a missingness proportion of 15% is used compared to 5%. The same cannot be said about the HMP however, as there does not appear to be a pattern here.

Finally, the number of indicator variables and the two loading options did not lead to any noticeable differences in the results, although the same pattern as in Situation 2 regarding the HMP and the loadings can be found again in both of our tables.

3.4. Situation 4 (the alternative hypothesis of 0.8 reflects the true situation)

In all of the combinations in Table A7 (describing the results of the datasets that were generated with missing data) and Table A8 (describing the results of the combined datasets), the power of the HMP is equal or higher than that of the mean p-value. The power of the mean p-value ranges from 0.639 to 1.000 (between 0.652 and 1.000 for the combined datasets) and the power of the HMP ranges between 0.882 and 1.000 (between 0.894 and 1.000 for the combined datasets). A substantial difference in the results can easily be found when one compares the results of the combinations where a sample size of 14 was used to the combinations which used 40 as a sample size. In the combinations with a sample size of 14, the power of the mean p-value ranges from 0.639 to 0.729 (between 0.652 and 0.741 in the combined datasets) whereas this range lies between 0.882 and 0.934 for the HMP (between 0.894 and 0.932 in the combined datasets). The power becomes higher in combinations where 40 was used as a sample size. In these combinations, the mean p-value ranges from 0.996 to 1.000 (from 0.997 to 1.000 in the combined datasets) and the HMP from 0.999 to 1.000 (also in the combined datasets). Notice how there are several combinations where the mean p-value and/or the HMP have a perfect result of 1.000, which means that a significant mean p-value and/or HMP was found in all the datasets for that combination. In a single pathway analysis, the power would be 0.8 for a sample size of 14 and 0.999 for a sample size of 40. In our results, we again see that with a sample size of 14, the power of the mean p-value is approximately 0.1 lower than that of a single pathway analysis in each of the combinations. However, the power of the mean p-value approximates that of a single pathway analysis rather well in the combinations which used a sample size of 40, since the power of the mean p-value ranges from 0.996 to 1.000 in these combinations (when we look at both tables). Meanwhile, the HMP is higher in power by approximately 0.1 compared to a single pathway analysis in the combinations with a sample size of 14. With a sample size of 40, the HMP is either equal to the single pathway in terms of power or higher with a perfect result of 1.000.

In addition, in Table A7 we can see that the number of datasets that were generated to have missing data differed quite a lot, as it ranges from 482 to 999. Around half of the datasets were generated to have missing data in the combinations with a sample size of 14 and a missingness proportion of 5%. This number is higher for the other combinations, where around 90% of all the datasets were generated with missing data. The number of observations is highest in the combinations where a sample size of 40 and a missingness rate of 15% was used, as the observations ranges from 997 to 999 in these combinations. When we again look

at the results of both tables, we can see that the combinations where a missingness proportion of 15% was used with a sample size of 14 are lower in power for the mean p-value compared to when a 5% missingness rate was used with this sample size. No pattern can be found in the combinations where a sample size of 40 was used, nor can any pattern regarding the missingness proportion be found for the HMP.

Finally, the number of indicator variables and the loadings lead to some slim differences in the results. While the aforementioned pattern of the HMP having more power in combinations where the loading was based on a medium correlation cannot consistently be found in Table A7, it can again be found in Table A8 (presenting the results of the combined datasets), but only in the combinations with a sample size of 14. Table A8 also shows us that in the combinations with a sample size of 14, the power of the mean p-value tends to go down if the loading was based on a medium correlation compared to if it was based on a large correlation. However, the differences are slim, seeing how the largest change is 0.050 for the mean p-value and 0.032 for the HMP.

Discussion

4.1. Concluding remarks about the summarization methods and performance of the multiverse analysis

In this simulation study, we examined the performance of the multiverse analysis using two summarization methods, specifically the mean p-value and the HMP. The results showed that the mean p-value was able to find a true negative result more often than the HMP. In addition, both summarization methods appeared to be able to find this result rather well, which is reflected through the relatively high proportion of finding a true negative result. While this does seem positive, it should be acknowledged that the HMP has a relatively high false-positive rate of almost 10% in some situations. Of course, falsely rejecting the null hypothesis is considered to be problematic when doing research. Therefore, one might consider the mean p-value to be the optimal method of more formally summarizing the results of a multiverse analysis in this situation, as it provides a more conservative result compared to the HMP.

When we examine situations where there is a real effect, we see that the HMP provides a significant result more often than the mean p-value. Thus, the HMP might be considered a superior summarization method compared to the mean p-value in these situations, as it more often reflects the expected result of showing significance. However, at this point it is important to discuss a large limitation regarding the summarization of the results of a multiverse analysis using either of these two methods. First, there is a risk that researchers would interpret the summarized result at face value and would not look at the results of the single pathways. By doing this, certain patterns or effects within the multiverse might be missed by the researcher. Even if the mean p-value or HMP returned a non-significant result, there might still be patterns within the multiverse that could be interesting to the researcher (and might warrant future research). In addition, finding a significant mean p-value or HMP does not necessarily mean that researchers can be sure that there is strong evidence for the existence of an overall effect. This is especially true when one uses the HMP, since it is more liberal in presenting a significant result. As was mentioned before, the HMP merely presents a significant result if at least one of the paths in the multiverse returns a significant result while correcting for multiple testing. Researchers should therefore always investigate the results of the different paths further and not only base conclusions on the result of the HMP (or mean p-value). Of course, determining whether the mean p-value or

HMP is the optimal or preferred summarization method also depends on the intent of the researcher. For example, if a researcher quickly wants to see if there is at least one path in the multiverse that is significant (i.e., shows an effect), the HMP would be the optimal choice since the mean p-value is not capable of providing this information. Essentially, one could think of the mean p-value and the HMP as two summarization methods that use two separate perspectives to summarize the results of a multiverse analysis. However, as this study has demonstrated, when there is no effect to be found in the data and a researcher uses the HMP for exactly this purpose (seeing whether there is an effect), the HMP would lead to a false-positive rate of around 10%. Thus, knowing that there is a risk of researchers interpreting the summarized result at face-value, and knowing that the HMP is more liberal in regards to returning a significant result which may lead to more false-positives, the mean p-value might actually be the preferred method of summarizing the results of a multiverse analysis.

Finally, some remarks about the different research manipulations can be made. Overall, the number of indicator variables, the proportion missingness options and the loading options appeared to have a minimal impact on the results of our generic multiverse. Essentially, one could say that the summarization methods used for our generic multiverse appear to be relatively robust against these variations. The same cannot be said about the sample size however. For example, this is most apparent in Situation 2 where a sample size of 40 is not nearly enough to find a small effect. As a result, the power of both the mean p-value and HMP becomes severely lower in these combinations compared to the combinations which use 198 as a sample size. Therefore, it is important to emphasize that researchers should use a large enough sample size if they wish to be able to find a small effect. While the HMP is still better at finding a significant small effect with small sample sizes compared to the mean p-value, it is not recommended to rely on the results of a multiverse analysis under these circumstances, since the analysis is not robust against (severely) underpowered data. Of course, this is true for every analysis and should therefore not be considered as a limitation that solely belongs to a multiverse analysis.

4.2. General limitations and future research

It is also important to mention some more general limitations regarding the results. First, because the proportion was always calculated from the total number of observations in each combination, the first table in each research situation (Table A1, A3, A5 and A7) that describe the results solely of the datasets that were generated with missing data, become

somewhat more difficult to interpret due to the difference in the number of observations for each combination. For example, Table A7 presents the results of the datasets that were generated with missing data. Here, some combinations have a fairly low number of observations (e.g., 482), while others have very high number of observations (e.g., 999). Accurately comparing these combinations with each other by calculating the proportion from their respective number of observations may be difficult because of the relatively large difference in sample size. For example, it might be possible that the results of some combinations with a low number of observations may have changed if more datasets were generated. This problem also arises if one were to compare the results of a combination with a low number of observations in Table A7 with the result of that same combination in Table A8. However, it should also be mentioned that several hundred observations for a certain combination is still a fair amount, which leads us to believe that any changes that could have occurred if more observations would have been generated would likely have had a limited impact on the results.

Second, it should again be mentioned that the results stemming from the datasets that were generated with missingness are somewhat underpowered due to the properties of listwise and pairwise deletion, and because of the (possible) removal of outliers. This could be (part of) the reason why the results of combinations which use 40 as a sample size appear to be higher in power than the results of the combinations which use 33 as a sample size in Table A6 of Situation 3 for example. Since the power calculation was performed without taking missing data into account, it is likely that a sample size of 33 was not enough to find a significant mean p-value or HMP in some instances. Thus, one should keep in mind that the power of the mean p-value and HMP of the combinations that are based on a power calculation are likely to be somewhat underestimated in the current study.

Finally, we did not take into account the possibility that no outliers were created in a specific dataset. If no outliers were created, some identical p-values would occur in some of the paths of the multiverse which could skew the results somewhat. Unfortunately, we are unable to say whether and if so, to what extent this had an effect on the results of the current study.

Some more general limitations about the current study can be named as well. For example, the create_data function is also able to simulate missingness that is MNAR. While this was outside the scope of the current study, it might be very interesting to investigate whether this would affect the results of a multiverse analysis and how the results would compare to our simulation study which simulated missing data that were MCAR. In addition,

the create_data function always simulates data from a model with two factors and users are able to set the correlation between the two factors. While it was outside of the scope of the current study, it might be interesting to examine how the multiverse analysis would perform in a correlational study. The multiverse_cor function in Appendix C is able to do this, and is shared so that other researchers may use this function if they want to investigate this further. Please note that if researchers want to use this function, they should first alter the create_data function so that missingness is generated across the whole dataset and not just on the items of the first factor.

In addition, some remarks can be made about the multiverse analysis and researcher degrees of freedom. While the multiverse analysis indeed limits the ability of authors to selectively use researcher degrees of freedom, being transparent about the chosen and performed analysis is still essential. Researchers are still able to selectively use researcher degrees of freedom if they, for example, only mention certain paths as being a part of the multiverse analysis and remove any mention of the paths that showed less favorable results. While ill-intent might not always be involved, we strongly urge researchers who perform a multiverse analysis to be transparent about the reasoning as to why certain paths were or were not included as paths in their multiverse. Pre-registration is also recommended, as this limits the ability of authors to selectively use researcher degrees of freedom.

However, it should be mentioned that the choices that we have made in the current study are somewhat subjective, which may lead to a problem regarding the generalizability of the results. While we have tried to be transparent about the reasoning behind our choices, there are of course things that we could have done differently. For example, our definition of a generic multiverse may very well be different from that of other researchers and fields of research. It may be possible that the number of indicator variables is generally much higher in other research fields. Or it might be possible that a specific technique for handling outliers (e.g., winsorization) would have been included in the generic multiverse by other researchers since they see that this is often used in their field. As a consequence, our results may have poor generalizability due to the variety of possible (generic) multiverse analyses that could be created by different researchers (in different research fields). Thus, we encourage other researchers to conduct future research on the performance of the multiverse analysis. This way, results can be compared. The functions that were used in the current study are shared in Appendix B. Other researchers are encouraged to expand on these functions further to use in their own studies.

In conclusion, we recommend using the mean p-value as a method to summarize the results of a multiverse analysis more formally, as it provides a more conservative result compared to the HMP and limits the chance of getting a false-positive result under the null hypothesis. In addition, this simulation study has demonstrated that our generic multiverse performs rather well, provided that the single pathways within the multiverse have adequate power.

References

Bakker, M., & Wicherts, J. M. (2014). Outlier removal, sum scores, and the inflation of the Type I error rate in independent samples t tests. The power of alternatives and recommendations. *Psychol. Methods* 19, 409–427. doi:10.1037/met0000014

Cesario, J., Johnson, D. J., & Terrill, W. (2018). Is There Evidence of Racial Disparity in Police Use of Deadly Force? Analyses of Officer-Involved Fatal Shootings in 2015–2016. *Social Psychological and Personality Science, 10*(5), doi:10.1177/1948550618775108

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erblaum Associates.

Dejonckheere, E., Mestdagh, M., Houben, M., Erbas, Y., Pe, M., Koval, P., . . . Kuppens, P. (2018). The Bipolarity of Affect and Depressive Symptoms. *Journal of Personality and Social Psychology, 114*(2), 323-341. doi:10.1037/pspp0000186

Denny, M. J., & Spirling, A. (2018). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Political Analysis, 26*(2), 168-189. doi:10.1017/pan.2017.44

Dong, Y., & Peng, C. Y. (2013). Principled missing data methods for researchers. *SpringerPlus, 2*(1), 222. doi:10.1186/2193-1801-2-222

Enders, C. K. (2003). Using the Expectation Maximization Algorithm to Estimate Coefficient Alpha for Scales With Item-Level Missing Data. *Psychological Methods, 8*(3), 322-337. doi:10.1037/1082-989X.8.3.322.

Champely, S. (2020). pwr: Basic Functions for Power Analysis [R package]. Retrieved from https://CRAN.R-project.org/package=pwr

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology, 64*(5), 402-406. doi:10.4097/kjae.2013.64.5.402

Liebst, L. S., Philpot, R., Bernasco, W., Dausel, K. L., Ejbye-ernst, P., Nicolaisen, M. H., & Lindegaard, M. R. (2019). Social relations and presence of others affect bystander

intervention: Evidence from violent incidents captured on CCTV. *Aggressive Behavior, 45*(6), 598–609. doi:10.1002/ab.21853

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual & Motor Skills, 112*(2), 331–348. doi:10.2466/03.11.pms .112.2.331-348

Pornprasertmanit, S., Miller, P., Schoemann, A., & Jorgensen, T.D. (2020). simsem: SIMulated Structural Equation Modeling [R package]. Retrieved from https://CRAN.R-project.org/package=simsem

Revelle, W. (2020). psych: Procedures for Personality and Psychological Research [R package]. Retrieved from https://CRAN.R-project.org/package=psych

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research, 8*(1), 3-15. doi: 10.1177/096228029900800102

Schafer, J. L. (2000). Missing Data: Our View of the State of the Art. *Psychological Methods, 7*(2), 147-177. doi: 10.1037//1082-989X.7.2.147

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359-1366. doi: 10.1177/0956797611417632

Simsem. (2021). Example 1: Getting started [Github wiki of simsem package]. Retrieved from https://github.com/simsem/simsem/wiki/Example-1:-Getting-Started

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11* (5), 702–712. doi:10.1177/1745691616658637

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, 45*(3), 1-67. doi:10.18637/jss.v045.i03

Wilson, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences U.S.A., 116*(4), 1195-1200. doi:10.1073/pnas.1814092116

Xie, Y. (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. [R package]. Retrieved from https://CRAN.R-project.org/package=knitr

Appendix A – Tables of the results

Table A1

Results of research Situation 1 (null hypothesis) with datasets that had missingness only

	proportion true negative of mean p-value	proportion true negative of harmonic mean p-value	number of observations
sample_size = 198 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.5	0.991	0.940	1000
sample_size = 198 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.3	0.998	0.934	1000
sample_size = 198 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.5	0.990	0.944	1000
sample_size = 198 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.3	0.995	0.938	1000
sample_size = 198 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.5	0.992	0.938	1000
sample_size = 198 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.3	0.997	0.938	1000
sample_size = 198 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.5	0.993	0.942	1000
sample_size = 198 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.3	0.997	0.941	1000
sample_size = 40 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.5	0.991	0.915	875
sample_size = 40 missing_prop = 0.05	0.999	0.927	852

nIndicator = 3 loadings = square root of 0.3			
sample_size = 40 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.5	0.997	0.931	872
sample_size = 40 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.3	0.998	0.936	895
sample_size = 40 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.5	0.993	0.918	999
sample_size = 40 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.3	0.996	0.922	999
sample_size = 40 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.5	0.993	0.928	998
sample_size = 40 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.3	0.999	0.924	997

Table A2

Combined results of research Situation 1 (null hypothesis) with the datasets that had both missingness and no missingness

	proportion true negative of mean p-value	proportion true negative of harmonic mean p-value	number of observations
sample_size = 198 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.5	0.991	0.940	1000
sample_size = 198 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.3	0.998	0.934	1000
sample_size = 198 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.5	0.990	0.944	1000

sample_size = 198 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.3	0.995	0.938	1000
sample_size = 198 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.5	0.992	0.938	1000
sample_size = 198 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.3	0.997	0.938	1000
sample_size = 198 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.5	0.993	0.942	1000
sample_size = 198 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.3	0.997	0.941	1000
sample_size = 40 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.5	0.992	0.919	1000
sample_size = 40 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.3	0.999	0.918	1000
sample_size = 40 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.5	0.997	0.931	1000
sample_size = 40 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.3	0.998	0.937	1000
sample_size = 40 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.5	0.993	0.918	1000
sample_size = 40 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.3	0.996	0.921	1000
sample_size = 40 missing_prop = 0.15	0.993	0.928	1000

nIndicator = 5 loadings = square root of 0.5			
sample_size = 40 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.3	0.999	0.924	1000

Table A3

Results of research Situation 2 (Cohen's d of 0.2) with datasets that had missingness only

	power of mean p-value	power of harmonic mean p-value	number of observations
sample_size = 198 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.5	0.728	0.890	1000
sample_size = 198 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.3	0.715	0.939	1000
sample_size = 198 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.5	0.736	0.907	1000
sample_size = 198 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.3	0.708	0.931	1000
sample_size = 198 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.5	0.689	0.896	1000
sample_size = 198 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.3	0.696	0.934	1000
sample_size = 198 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.5	0.710	0.892	1000
sample_size = 198 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.3	0.726	0.939	1000
sample_size = 40 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.5	0.125	0.323	875
sample_size = 40 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.3	0.085	0.345	852
sample_size = 40 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.5	0.101	0.298	872

sample_size = 40 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.3	0.088	0.323	895
sample_size = 40 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.5	0.106	0.325	999
sample_size = 40 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.3	0.082	0.365	999
sample_size = 40 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.5	0.088	0.323	998
sample_size = 40 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.3	0.069	0.328	997

Table A4

Combined results of research Situation 2 (Cohen's d of 0.2) with the datasets that had both missingness and no missingness

	power of mean p-value	power of harmonic mean p-value	number of observations
sample_size = 198 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.5	0.728	0.890	1000
sample_size = 198 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.3	0.715	0.939	1000
sample_size = 198 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.5	0.736	0.907	1000
sample_size = 198 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.3	0.708	0.931	1000
sample_size = 198 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.5	0.689	0.896	1000
sample_size = 198 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.3	0.696	0.934	1000
sample_size = 198 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.5	0.710	0.892	1000

sample_size = 198 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.3	0.726	0.939	1000
sample_size = 40 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.5	0.123	0.318	1000
sample_size = 40 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.3	0.090	0.355	1000
sample_size = 40 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.5	0.106	0.302	1000
sample_size = 40 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.3	0.087	0.320	1000
sample_size = 40 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.5	0.106	0.326	1000
sample_size = 40 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.3	0.083	0.366	1000
sample_size = 40 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.5	0.088	0.322	1000
sample_size = 40 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.3	0.069	0.328	1000

Table A5

Results of research Situation 3 (Cohen's d of 0.5) with datasets that had missingness only

	power of mean p-value	power of harmonic mean p-value	number of observations
sample_size = 33 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.5	0.735	0.897	815
sample_size = 33 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.3	0.691	0.923	832
sample_size = 33 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.5	0.730	0.880	830
sample_size = 33 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.3	0.734	0.930	811

sample_size = 33 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.5	0.704	0.900	990
sample_size = 33 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.3	0.690	0.938	995
sample_size = 33 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.5	0.700	0.886	992
sample_size = 33 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.3	0.696	0.932	997
sample_size = 40 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.5	0.811	0.938	875
sample_size = 40 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.3	0.812	0.980	852
sample_size = 40 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.5	0.838	0.945	872
sample_size = 40 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.3	0.854	0.980	895
sample_size = 40 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.5	0.811	0.935	999
sample_size = 40 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.3	0.812	0.971	999
sample_size = 40 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.5	0.821	0.955	998
sample_size = 40 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.3	0.803	0.963	997

Table A6

Combined results of research Situation 3 (Cohen's d of 0.5) with the datasets that had both missingness and no missingness

	power of mean p-value	power of harmonic mean p-value	number of observations

sample_size = 33 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.5	0.747	0.910	1000
sample_size = 33 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.3	0.705	0.930	1000
sample_size = 33 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.5	0.737	0.889	1000
sample_size = 33 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.3	0.736	0.936	1000
sample_size = 33 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.5	0.706	0.901	1000
sample_size = 33 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.3	0.692	0.938	1000
sample_size = 33 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.5	0.701	0.886	1000
sample_size = 33 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.3	0.695	0.931	1000
sample_size = 40 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.5	0.814	0.936	1000
sample_size = 40 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.3	0.819	0.982	1000
sample_size = 40 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.5	0.842	0.945	1000
sample_size = 40 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.3	0.848	0.976	1000
sample_size = 40 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.5	0.811	0.935	1000
sample_size = 40 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.3	0.812	0.971	1000
sample_size = 40 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.5	0.819	0.954	1000

sample_size = 40 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.3	0.802	0.962	1000

Table A7

Results of research Situation 4 (Cohen's d of 0.8) with datasets that had missingness only

	power of mean p-value	power of harmonic mean p-value	number of observations
sample_size = 14 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.5	0.707	0.882	482
sample_size = 14 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.3	0.661	0.934	531
sample_size = 14 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.5	0.723	0.896	531
sample_size = 14 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.3	0.729	0.931	494
sample_size = 14 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.5	0.698	0.906	896
sample_size = 14 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.3	0.639	0.927	894
sample_size = 14 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.5	0.677	0.893	897
sample_size = 14 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.3	0.671	0.921	911
sample_size = 40 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.5	1.000	1.000	875
sample_size = 40 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.3	0.996	0.999	852
sample_size = 40 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.5	1.000	1.000	872
sample_size = 40 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.3	0.999	1.000	895

sample_size = 40 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.5	0.998	1.000	999
sample_size = 40 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.3	0.998	1.000	999
sample_size = 40 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.5	0.999	1.000	998
sample_size = 40 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.3	0.998	1.000	997

Table A8

Combined results of research Situation 4 (Cohen's d of 0.8) with the datasets that had both missingness and no missingness

	power of mean p-value	power of harmonic mean p-value	number of observations
sample_size = 14 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.5	0.738	0.900	1000
sample_size = 14 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.3	0.712	0.932	1000
sample_size = 14 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.5	0.741	0.899	1000
sample_size = 14 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.3	0.734	0.931	1000
sample_size = 14 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.5	0.702	0.905	1000
sample_size = 14 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.3	0.652	0.932	1000
sample_size = 14 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.5	0.685	0.894	1000
sample_size = 14 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.3	0.681	0.922	1000

sample_size = 40 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.5	1.000	1.000	1000
sample_size = 40 missing_prop = 0.05 nIndicator = 3 loadings = square root of 0.3	0.997	0.999	1000
sample_size = 40 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.5	1.000	1.000	1000
sample_size = 40 missing_prop = 0.05 nIndicator = 5 loadings = square root of 0.3	0.999	1.000	1000
sample_size = 40 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.5	0.998	1.000	1000
sample_size = 40 missing_prop = 0.15 nIndicator = 3 loadings = square root of 0.3	0.998	1.000	1000
sample_size = 40 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.5	0.999	1.000	1000
sample_size = 40 missing_prop = 0.15 nIndicator = 5 loadings = square root of 0.3	0.998	1.000	1000

Appendix B – R code

Load packages

````{r load libraries, message = F, warning = F}
library(simsem)
library(psych)
library(mice)
library(harmonicmeanp)
library(pwr)
library(knitr)
```

# Load functions

```{r tool functions}
# Create data function:

create_data <- function(sample_size = 200, nFactor = 2, nIndicator = 3, correlation = 0.5,
indicator_means = 0.2, loadings = sqrt(0.5), missing_type = "MCAR", missing_prop = 0.05,
seed = runif(n = 1, min = 0, max = 10000)){
  loading <- matrix(0, nIndicator * nFactor, nFactor)
  loading[1:nIndicator, 1] <- NA
  loading[(1 + nIndicator):(nIndicator * nFactor), 2] <- NA

  loadingValues <- matrix(0, nIndicator * nFactor, nFactor)
  loadingValues[1:nIndicator, 1] <- loadings
  loadingValues[(1 + nIndicator):(nIndicator * nFactor), 2] <- loadings

  # LY = factor loading matrix
  LY <- bind(loading, loadingValues)

  # RTE = error correlation matrix
  error.cor <- matrix(0, nIndicator * nFactor, nIndicator * nFactor)
  diag(error.cor) <- 1
```

```
  RTE <- binds(error.cor)

  # RPS = factor correlation matrix
  latent.cor <- matrix(NA, nFactor, nFactor)
  diag(latent.cor) <- 1
  RPS <- binds(latent.cor, correlation)

  # Set indicator means and create CFA model;
  ind.mean <- rep(NA, nIndicator * nFactor)
  ind.mean.starting <- c(rep(indicator_means, nIndicator * nFactor))
  AL <- bind(ind.mean, ind.mean.starting)
  CFA.Model <- model(LY = LY, RPS = RPS, RTE = RTE, modelType="CFA", MY=AL)

  # Create data
  set.seed(seed)
  dat <- generate(CFA.Model,sample_size)
  if (missing_type == "MCAR"){
    set.seed(seed)
    mcar_dat <- ampute(data = dat[,1:nIndicator], mech = "MCAR", prop = missing_prop)
    dat <- mcar_dat$amp
  }
  if (missing_type == "MNAR"){
    set.seed(seed)
    mnar_dat <- ampute(data = dat[,1:nIndicator], mech = "MNAR", prop = missing_prop)
    dat <- mnar_dat$amp
  }
  if (missing_type != "MCAR" & missing_type != "MNAR"){
    warning("Invalid missing_type. No missing data was generated")
  }
  return(dat)
}
#------------------------------------------------------------------------------

# T-test multiverse function:
```

```
multiverse_ttest <- function(nFactor = 2, nIndicator = 3, method_missing = c("listwise",
"pairwise", "mi"), method_outliers = c("none", "IQR_remove", "z_remove"), dat =
create_data()){
  result = list()
  for (i in 1:12){
    result[[i]] <- NA
  }
  if (sum(method_missing %in% c("listwise", "pairwise","mi") & "none" %in%
method_missing ) >= 1){
    stop("Invalid combination")
  }
  if ("IQR_remove" %in% method_outliers){
    IQR_dat <- dat[, 1:nIndicator]
    for (i in 1:nIndicator) {
      IQR_dat[, i][IQR_dat[, i] > summary(IQR_dat[, i])[5] + 1.5 * IQR(IQR_dat[, i], na.rm =
T) | IQR_dat[, i] < summary(IQR_dat[, i])[2] -1.5 * IQR(IQR_dat[, i], na.rm = T)] <- NA
    }
  }
  if("z_remove" %in% method_outliers){
    z_dat <- dat[, 1:nIndicator]
    scaled_dat <- scale(z_dat)
    for (i in 1:nIndicator){
      z_dat[, i][scaled_dat[, i] > 3 | scaled_dat[, i] < -3] <- NA
    }
  }
  if ("none" %in% method_missing){
    if (sum(is.na(dat)) != 0){
      warning("The data has missing values")
    }
  }
  if ("none" %in% method_missing & "none" %in% method_outliers){
    for (i in 1:nIndicator){
      result[[1]][i] <- t.test(dat[, i])$p.value
    }
```

```
}
if ("none" %in% method_missing & "IQR_remove" %in% method_outliers){
  for (i in 1:nIndicator){
    result[[2]][i] <- t.test(IQR_dat[, i])$p.value
  }
}
if ("none" %in% method_missing & "z_remove" %in% method_outliers){
  for (i in 1:nIndicator){
    result[[3]][i] <- t.test(z_dat[, i])$p.value
  }
}
if ("listwise" %in% method_missing & "none" %in% method_outliers){
  temp_dat <- dat[, 1:nIndicator]
  complete_dat <- temp_dat[complete.cases(temp_dat),]
  for (i in 1:nIndicator){
    result[[4]][i] <- t.test(complete_dat[, i])$p.value
  }
}
if ("listwise" %in% method_missing & "IQR_remove" %in% method_outliers){
  IQR_complete_dat <- IQR_dat[complete.cases(IQR_dat),]
  for (i in 1:nIndicator){
    result[[5]][i] <- t.test(IQR_complete_dat[, i])$p.value
  }
}
if ("listwise" %in% method_missing & "z_remove" %in% method_outliers){
  z_complete_dat <- z_dat[complete.cases(z_dat),]
  for (i in 1:nIndicator){
    result[[6]][i] <- t.test(z_complete_dat[, i])$p.value
  }
}
if ("pairwise" %in% method_missing & "none" %in% method_outliers){
  for (i in 1:nIndicator){
    result[[7]][i] <- t.test(dat[, i])$p.value
  }
```

```
  }
  if ("pairwise" %in% method_missing & "IQR_remove" %in% method_outliers){
    for (i in 1:nIndicator){
      result[[8]][i] <- t.test(IQR_dat[, i])$p.value
    }
  }
  if ("pairwise" %in% method_missing & "z_remove" %in% method_outliers){
    for (i in 1:nIndicator){
      result[[9]][i] <- t.test(z_dat[, i])$p.value
    }
  }
  if ("mi" %in% method_missing & "none" %in% method_outliers){
    capture.output(imputed <- mice(data = dat[, 1: nIndicator], seed = 2346982))
    df.imp <- complete(imputed)
    for (i in 1:nIndicator){
      result[[10]][i] <- t.test(df.imp[, i])$p.value
    }
  }
  if ("mi" %in% method_missing & "IQR_remove" %in% method_outliers){
    capture.output(IQR_imputed <- mice(data = IQR_dat, seed = 2346982))
    IQR_df.imp <- complete(IQR_imputed)
    for (i in 1:nIndicator){
      result[[11]][i] <- t.test(IQR_df.imp[, i])$p.value
    }
  }
  if("mi" %in% method_missing & "z_remove" %in% method_outliers){
    capture.output(z_imputed <- mice(data = z_dat, seed = 2346982))
    z_df.imp <- complete(z_imputed)
    for (i in 1:nIndicator){
      result[[12]][i] <- t.test(z_df.imp[, i])$p.value
    }
  }
  if (sum(method_missing %in% c("listwise", "pairwise","mi")) >= 1 & sum(is.na(dat)) ==
0){
```

```
    warning("The data has no missing values")
  }
  result_tot <- do.call(rbind, result)
  output <- matrix(NA, ncol = 2)
  output[,1] <- mean(result_tot, na.rm = T) # mean p-value
  output[,2] <- p.hmp(result_tot[! is.na(result_tot)],L=length(result_tot[! is.na(result_tot)]),
multilevel=FALSE) # harmonic mean p-value
  return(output)
}
```

# Default options

```{r default options}
# Default parameters for create_data:
# - Sample size = 200
# - 2 factors
# - 3 indicator variables per factor
# - Correlation of 0.5
# - Indicator mean of 0.2
# - Missingness that is MCAR
# - Proportion of 5% missingness

# Default in multiverse:
# - 2 factors
# - 3 indicator variables
# - All options for handling missing data
# - All options for handling outliers
# - Resulting in 3 (items) x 3 (options for handling missing data) x 3 (options for handling outliers) = 27 p-values that are used to calculate the mean p-value and harmonic mean p-value.
```

# Situation 1: Variations of the null hypothesis with missing data

````{r situation 1}
````

```
# Options/variations under the null hypothesis
options_sample_size <- c(198, 40)
options_missing_prop <- c(0.05, 0.15)
options_nIndicator <- c(3, 5)
options_loadings = c(sqrt(0.5), sqrt(0.3))

# Define n
n = 1000



# Get the results
pvalues_nomissing <- list()
pvalues_withmissing <- list()
overview_nomissing <- list()
overview_withmissing <- list()
iter <- 1
start <- proc.time()
for (i in 1:n){
  for (i_sample_size in 1:length(options_sample_size)){
    for (i_missing in 1:length(options_missing_prop)){
      for (i_indicators in 1:length(options_nIndicator)){
        for (i_loadings in 1:length(options_loadings)){
          dat <- create_data(indicator_means = 0, sample_size =
options_sample_size[i_sample_size], missing_prop = options_missing_prop [i_missing],
nIndicator = options_nIndicator [i_indicators], loadings = options_loadings [i_loadings], seed
= iter)
          if (sum(is.na(dat))==0){
            pvalues_nomissing[[iter]] <- multiverse_ttest(dat = dat, method_missing = 'none')
            overview_nomissing[[iter]] <- do.call(cbind, list(pvalues_nomissing[[iter]],
options_sample_size[[i_sample_size]], options_missing_prop[[i_missing]],
options_nIndicator[[i_indicators]], options_loadings[[i_loadings]]))
          }
          else{
            pvalues_withmissing[[iter]] <- multiverse_ttest(dat = dat)
```

```
        overview_withmissing[[iter]] <- do.call(cbind, list(pvalues_withmissing[[iter]],
options_sample_size[[i_sample_size]], options_missing_prop[[i_missing]],
options_nIndicator[[i_indicators]], options_loadings[[i_loadings]]))
        }
        iter <- iter + 1
    }
   }
  }
}
}
end <- proc.time()
end - start # find out how long it takes for the function to run
results_nomissing <- do.call(rbind, overview_nomissing)
colnames(results_nomissing) <- c("mean p-value", "harmonic p-value", "sample_size",
"missing_prop", "nIndicator", "loadings")
results_withmissing <- do.call(rbind, overview_withmissing)
colnames(results_withmissing) <- c("mean p-value", "harmonic p-value", "sample_size",
"missing_prop", "nIndicator", "loadings")


# Check for unique values and combine datasets
length(unique(results_nomissing[,1]))
length(unique(results_withmissing[,1]))
results_combined <- do.call(rbind, list(results_nomissing, results_withmissing))



# Calculate power (proportion that is correctly classified) and put it in a table
# Table of results_withmissing
results_withmissing <- as.data.frame(results_withmissing)
iter = 1
tab_results_withmissing <- matrix(NA, nrow = 16, ncol = 3)
names <- matrix(NA, nrow = 16, ncol = 1)
```

```
colnames(tab_results_withmissing) <- c("power of mean p-value", "power of harmonic mean p-value", "number of observations")
for (i in 1:length(options_sample_size)){
  for (j in 1: length(options_missing_prop)){
    for (k in 1: length(options_nIndicator)){
      for (l in 1:length(options_loadings)){
        tab_results_withmissing[iter, 1] <- sum(results_withmissing[results_withmissing$sample_size == options_sample_size[i] & results_withmissing$missing_prop == options_missing_prop[j] & results_withmissing$nIndicator == options_nIndicator[k] & results_withmissing$loadings == options_loadings[l], 1] > 0.05)/length(results_withmissing[results_withmissing$sample_size == options_sample_size[i] & results_withmissing$missing_prop == options_missing_prop[j] & results_withmissing$nIndicator == options_nIndicator[k] & results_withmissing$loadings == options_loadings[l], 1])
        tab_results_withmissing[iter, 2] <- sum(results_withmissing[results_withmissing$sample_size == options_sample_size[i] & results_withmissing$missing_prop == options_missing_prop[j] & results_withmissing$nIndicator == options_nIndicator[k] & results_withmissing$loadings == options_loadings[l], 2] > 0.05)/length(results_withmissing[results_withmissing$sample_size == options_sample_size[i] & results_withmissing$missing_prop == options_missing_prop[j] & results_withmissing$nIndicator == options_nIndicator[k] & results_withmissing$loadings == options_loadings[l], 2])
        tab_results_withmissing[iter, 3] <- length(results_withmissing[results_withmissing$sample_size == options_sample_size[i] & results_withmissing$missing_prop == options_missing_prop[j] & results_withmissing$nIndicator == options_nIndicator[k] & results_withmissing$loadings == options_loadings[l], 1]) # is the same number for column 2
        names[iter] <- c(paste0("sample_size = ", options_sample_size[i], " missing_prop = ", options_missing_prop[j], " nIndicator = ", options_nIndicator[k], " loadings = ", options_loadings[l]))
        iter = iter + 1
      }
    }
  }
```

```
}
rownames(tab_results_withmissing) <- names
kable(tab_results_withmissing)

# Table of results_combined
results_combined <- as.data.frame(results_combined)
iter = 1
tab_results_combined <- matrix(NA, nrow = 16, ncol = 3)
names <- matrix(NA, nrow = 16, ncol = 1)
colnames(tab_results_combined) <- c("power of mean p-value", "power of harmonic mean p-value", "number of observations")
for (i in 1:length(options_sample_size)){
  for (j in 1: length(options_missing_prop)){
    for (k in 1: length(options_nIndicator)){
      for (l in 1:length(options_loadings)){
        tab_results_combined[iter, 1] <- sum(results_combined[results_combined$sample_size == options_sample_size[i] & results_combined$missing_prop == options_missing_prop[j] & results_combined$nIndicator == options_nIndicator[k] & results_combined$loadings == options_loadings[l], 1] > 0.05)/n
        tab_results_combined[iter, 2] <- sum(results_combined[results_combined$sample_size == options_sample_size[i] & results_combined$missing_prop == options_missing_prop[j] & results_combined$nIndicator == options_nIndicator[k] & results_combined$loadings == options_loadings[l], 2] > 0.05)/n
        tab_results_combined[iter, 3] <- length(results_combined[results_combined$sample_size == options_sample_size[i] & results_combined$missing_prop == options_missing_prop[j] & results_combined$nIndicator == options_nIndicator[k] & results_combined$loadings == options_loadings[l], 1]) # is the same number for column 2
        names[iter] <- c(paste0("sample_size = ", options_sample_size[i], " missing_prop = ", options_missing_prop[j], " nIndicator = ", options_nIndicator[k], " loadings = ", options_loadings[l]))
        iter = iter + 1
      }
    }
```

```
  }
}
rownames(tab_results_combined) <- names
kable(tab_results_combined)
```

Situation 2: Variations of the altenative hypothesis (Ha = 0.2) with missing data

```{r situation 2}
# Power analysis to find sample_size
pwr.t.test(d = 0.2, power = 0.8, type = "one.sample")
# sample size should be 198 (rounded off)


# Options/variations under the alternative hypothesis of 0.2
options_sample_size <- c(198, 40)
options_missing_prop <- c(0.05, 0.15)
options_nIndicator <- c(3, 5)
options_loadings = c(sqrt(0.5), sqrt(0.3))


# Get the results
pvalues_nomissing <- list()
pvalues_withmissing <- list()
overview_nomissing <- list()
overview_withmissing <- list()
iter <- 1
start <- proc.time()
for (i in 1:n){
  for (i_sample_size in 1:length(options_sample_size)){
    for (i_missing in 1:length(options_missing_prop)){
      for (i_indicators in 1:length(options_nIndicator)){
        for (i_loadings in 1:length(options_loadings)){
```

```
        dat <- create_data(sample_size = options_sample_size[i_sample_size], missing_prop =
options_missing_prop [i_missing], nIndicator = options_nIndicator [i_indicators], loadings =
options_loadings [i_loadings], seed = iter)# note that indicator_means = 0.2 (default)
        if (sum(is.na(dat))==0){
          pvalues_nomissing[[iter]] <- multiverse_ttest(dat = dat, method_missing = 'none')
          overview_nomissing[[iter]] <- do.call(cbind, list(pvalues_nomissing[[iter]],
options_sample_size[[i_sample_size]], options_missing_prop[[i_missing]],
options_nIndicator[[i_indicators]], options_loadings[[i_loadings]]))
        }
        else{
          pvalues_withmissing[[iter]] <- multiverse_ttest(dat = dat)
          overview_withmissing[[iter]] <- do.call(cbind, list(pvalues_withmissing[[iter]],
options_sample_size[[i_sample_size]], options_missing_prop[[i_missing]],
options_nIndicator[[i_indicators]], options_loadings[[i_loadings]]))
        }
        iter <- iter + 1
      }
    }
  }
}
}
end <- proc.time()
end - start
results_nomissing <- do.call(rbind, overview_nomissing)
colnames(results_nomissing) <- c("mean p-value", "harmonic p-value", "sample_size",
"missing_prop", "nIndicator", "loadings")
results_withmissing <- do.call(rbind, overview_withmissing)
colnames(results_withmissing) <- c("mean p-value", "harmonic p-value", "sample_size",
"missing_prop", "nIndicator", "loadings")


# Check for unique values and combine datasets
length(unique(results_nomissing[,1]))
length(unique(results_withmissing[,1]))
```

```
results_combined <- do.call(rbind, list(results_nomissing, results_withmissing))



# Calculate power (proportion that is correctly classified) and put it in a table
# Table of results_withmissing
results_withmissing <- as.data.frame(results_withmissing)
iter = 1
tab_results_withmissing <- matrix(NA, nrow = 16, ncol = 3)
names <- matrix(NA, nrow = 16, ncol = 1)
colnames(tab_results_withmissing) <- c("power of mean p-value", "power of harmonic mean
p-value", "number of observations")
for (i in 1:length(options_sample_size)){
  for (j in 1: length(options_missing_prop)){
    for (k in 1: length(options_nIndicator)){
      for (l in 1:length(options_loadings)){
        tab_results_withmissing[iter, 1] <-
sum(results_withmissing[results_withmissing$sample_size == options_sample_size[i] &
results_withmissing$missing_prop == options_missing_prop[j] &
results_withmissing$nIndicator == options_nIndicator[k] & results_withmissing$loadings ==
options_loadings[l], 1] <=
0.05)/length(results_withmissing[results_withmissing$sample_size ==
options_sample_size[i] & results_withmissing$missing_prop == options_missing_prop[j] &
results_withmissing$nIndicator == options_nIndicator[k] & results_withmissing$loadings ==
options_loadings[l], 1])
        tab_results_withmissing[iter, 2] <-
sum(results_withmissing[results_withmissing$sample_size == options_sample_size[i] &
results_withmissing$missing_prop == options_missing_prop[j] &
results_withmissing$nIndicator == options_nIndicator[k] & results_withmissing$loadings ==
options_loadings[l], 2] <=
0.05)/length(results_withmissing[results_withmissing$sample_size ==
options_sample_size[i] & results_withmissing$missing_prop == options_missing_prop[j] &
results_withmissing$nIndicator == options_nIndicator[k] & results_withmissing$loadings ==
options_loadings[l], 2])
```

```
      tab_results_withmissing[iter, 3] <-
length(results_withmissing[results_withmissing$sample_size == options_sample_size[i] &
results_withmissing$missing_prop == options_missing_prop[j] &
results_withmissing$nIndicator == options_nIndicator[k] & results_withmissing$loadings ==
options_loadings[l], 1])
      names[iter] <- c(paste0("sample_size = ", options_sample_size[i], " missing_prop = ",
options_missing_prop[j], " nIndicator = ", options_nIndicator[k], " loadings = ",
options_loadings[l]))
      iter = iter + 1
    }
   }
  }
}
rownames(tab_results_withmissing) <- names
kable(tab_results_withmissing)

# Table of results_combined
results_combined <- as.data.frame(results_combined)
iter = 1
tab_results_combined <- matrix(NA, nrow = 16, ncol = 3)
names <- matrix(NA, nrow = 16, ncol = 1)
colnames(tab_results_combined) <- c("power of mean p-value", "power of harmonic mean p-
value", "number of observations")
for (i in 1:length(options_sample_size)){
  for (j in 1: length(options_missing_prop)){
    for (k in 1: length(options_nIndicator)){
      for (l in 1:length(options_loadings)){
      tab_results_combined[iter, 1] <- sum(results_combined[results_combined$sample_size
== options_sample_size[i] & results_combined$missing_prop == options_missing_prop[j] &
results_combined$nIndicator == options_nIndicator[k] & results_combined$loadings ==
options_loadings[l], 1] <= 0.05)/n
      tab_results_combined[iter, 2] <- sum(results_combined[results_combined$sample_size
== options_sample_size[i] & results_combined$missing_prop == options_missing_prop[j] &
```

```
results_combined$nIndicator == options_nIndicator[k] & results_combined$loadings == options_loadings[l], 2] <= 0.05)/n
        tab_results_combined[iter, 3] <- length(results_combined[results_combined$sample_size == options_sample_size[i] & results_combined$missing_prop == options_missing_prop[j] & results_combined$nIndicator == options_nIndicator[k] & results_combined$loadings == options_loadings[l], 1])
        names[iter] <- c(paste0("sample_size = ", options_sample_size[i], " missing_prop = ", options_missing_prop[j], " nIndicator = ", options_nIndicator[k], " loadings = ", options_loadings[l]))
        iter = iter + 1
      }
    }
  }
}
rownames(tab_results_combined) <- names
kable(tab_results_combined)
```

Situation 3: Variations of the altenative hypothesis (Ha = 0.5) with missing data

```{r situation 3}
# Power analysis to find sample_size
pwr.t.test(d = 0.5, power = 0.8, type = "one.sample")
# sample size should be 33 (rounded off)

# Options/variations under the alternative hypothesis of 0.2
options_sample_size <- c(33, 40)
options_missing_prop <- c(0.05, 0.15)
options_nIndicator <- c(3, 5)
options_loadings = c(sqrt(0.5), sqrt(0.3))


# Get the results
pvalues_nomissing <- list()
pvalues_withmissing <- list()
```

```
overview_nomissing <- list()
overview_withmissing <- list()
iter <- 1
start <- proc.time()
for (i in 1:n){
  for (i_sample_size in 1:length(options_sample_size)){
    for (i_missing in 1:length(options_missing_prop)){
      for (i_indicators in 1:length(options_nIndicator)){
        for (i_loadings in 1:length(options_loadings)){
          dat <- create_data(indicator_means = 0.5, sample_size =
options_sample_size[i_sample_size], missing_prop = options_missing_prop [i_missing],
nIndicator = options_nIndicator [i_indicators], loadings = options_loadings [i_loadings], seed
= iter)
          if (sum(is.na(dat))==0){
            pvalues_nomissing[[iter]] <- multiverse_ttest(dat = dat, method_missing = 'none')
            overview_nomissing[[iter]] <- do.call(cbind, list(pvalues_nomissing[[iter]],
options_sample_size[[i_sample_size]], options_missing_prop[[i_missing]],
options_nIndicator[[i_indicators]], options_loadings[[i_loadings]]))
          }
          else{
            pvalues_withmissing[[iter]] <- multiverse_ttest(dat = dat)
            overview_withmissing[[iter]] <- do.call(cbind, list(pvalues_withmissing[[iter]],
options_sample_size[[i_sample_size]], options_missing_prop[[i_missing]],
options_nIndicator[[i_indicators]], options_loadings[[i_loadings]]))
          }
          iter <- iter + 1
      }
    }
  }
}
}
end <- proc.time()
end - start
results_nomissing <- do.call(rbind, overview_nomissing)
```

```
colnames(results_nomissing) <- c("mean p-value", "harmonic p-value", "sample_size",
"missing_prop", "nIndicator", "loadings")
results_withmissing <- do.call(rbind, overview_withmissing)
colnames(results_withmissing) <- c("mean p-value", "harmonic p-value", "sample_size",
"missing_prop", "nIndicator", "loadings")


# Check for unique values and combine datasets
length(unique(results_nomissing[,1]))
length(unique(results_withmissing[,1]))
results_combined <- do.call(rbind, list(results_nomissing, results_withmissing))



# Put the results in a table (power)
# Table of results_withmissing
results_withmissing <- as.data.frame(results_withmissing)
iter = 1
tab_results_withmissing <- matrix(NA, nrow = 16, ncol = 3)
names <- matrix(NA, nrow = 16, ncol = 1)
colnames(tab_results_withmissing) <- c("power of mean p-value", "power of harmonic mean
p-value", "number of observations")
for (i in 1:length(options_sample_size)){
  for (j in 1: length(options_missing_prop)){
    for (k in 1: length(options_nIndicator)){
      for (l in 1:length(options_loadings)){
        tab_results_withmissing[iter, 1] <-
sum(results_withmissing[results_withmissing$sample_size == options_sample_size[i] &
results_withmissing$missing_prop == options_missing_prop[j] &
results_withmissing$nIndicator == options_nIndicator[k] & results_withmissing$loadings ==
options_loadings[l], 1] <=
0.05)/length(results_withmissing[results_withmissing$sample_size ==
options_sample_size[i] & results_withmissing$missing_prop == options_missing_prop[j] &
```

```
results_withmissing$nIndicator == options_nIndicator[k] & results_withmissing$loadings ==
options_loadings[l], 1])
      tab_results_withmissing[iter, 2] <-
sum(results_withmissing[results_withmissing$sample_size == options_sample_size[i] &
results_withmissing$missing_prop == options_missing_prop[j] &
results_withmissing$nIndicator == options_nIndicator[k] & results_withmissing$loadings ==
options_loadings[l], 2] <=
0.05)/length(results_withmissing[results_withmissing$sample_size ==
options_sample_size[i] & results_withmissing$missing_prop == options_missing_prop[j] &
results_withmissing$nIndicator == options_nIndicator[k] & results_withmissing$loadings ==
options_loadings[l], 2])
      tab_results_withmissing[iter, 3] <-
length(results_withmissing[results_withmissing$sample_size == options_sample_size[i] &
results_withmissing$missing_prop == options_missing_prop[j] &
results_withmissing$nIndicator == options_nIndicator[k] & results_withmissing$loadings ==
options_loadings[l], 1])
      names[iter] <- c(paste0("sample_size = ", options_sample_size[i], " missing_prop = ",
options_missing_prop[j], " nIndicator = ", options_nIndicator[k], " loadings = ",
options_loadings[l]))
      iter = iter + 1
    }
  }
 }
}
rownames(tab_results_withmissing) <- names
kable(tab_results_withmissing)

# Table of results_combined
results_combined <- as.data.frame(results_combined)
iter = 1
tab_results_combined <- matrix(NA, nrow = 16, ncol = 3)
names <- matrix(NA, nrow = 16, ncol = 1)
colnames(tab_results_combined) <- c("power of mean p-value", "power of harmonic mean p-
value", "number of observations")
```

```
for (i in 1:length(options_sample_size)){
  for (j in 1: length(options_missing_prop)){
    for (k in 1: length(options_nIndicator)){
      for (l in 1:length(options_loadings)){
        tab_results_combined[iter, 1] <- sum(results_combined[results_combined$sample_size == options_sample_size[i] & results_combined$missing_prop == options_missing_prop[j] & results_combined$nIndicator == options_nIndicator[k] & results_combined$loadings == options_loadings[l], 1] <= 0.05)/n
        tab_results_combined[iter, 2] <- sum(results_combined[results_combined$sample_size == options_sample_size[i] & results_combined$missing_prop == options_missing_prop[j] & results_combined$nIndicator == options_nIndicator[k] & results_combined$loadings == options_loadings[l], 2] <= 0.05)/n
        tab_results_combined[iter, 3] <- length(results_combined[results_combined$sample_size == options_sample_size[i] & results_combined$missing_prop == options_missing_prop[j] & results_combined$nIndicator == options_nIndicator[k] & results_combined$loadings == options_loadings[l], 1])
        names[iter] <- c(paste0("sample_size = ", options_sample_size[i], " missing_prop = ", options_missing_prop[j], " nIndicator = ", options_nIndicator[k], " loadings = ", options_loadings[l]))
        iter = iter + 1
      }
    }
  }
}
rownames(tab_results_combined) <- names
kable(tab_results_combined)
```

Situation 4: Variations of the altenative hypothesis (Ha = 0.8) with missing data

```{r situation 4}
# Power analysis to find sample_size
pwr.t.test(d = 0.8, power = 0.8, type = "one.sample")
# sample size should be 14 (rounded off)
```

```
# Options/variations under the alternative hypothesis of 0.2
options_sample_size <- c(14, 40)
options_missing_prop <- c(0.05, 0.15)
options_nIndicator <- c(3, 5)
options_loadings = c(sqrt(0.5), sqrt(0.3))

# Get the results
pvalues_nomissing <- list()
pvalues_withmissing <- list()
overview_nomissing <- list()
overview_withmissing <- list()
iter <- 1
start <- proc.time()
for (i in 1:n){
  for (i_sample_size in 1:length(options_sample_size)){
    for (i_missing in 1:length(options_missing_prop)){
      for (i_indicators in 1:length(options_nIndicator)){
        for (i_loadings in 1:length(options_loadings)){
          dat <- create_data(indicator_means = 0.8, sample_size =
options_sample_size[i_sample_size], missing_prop = options_missing_prop [i_missing],
nIndicator = options_nIndicator [i_indicators], loadings = options_loadings [i_loadings], seed
= iter)
          if (sum(is.na(dat))==0){
            pvalues_nomissing[[iter]] <- multiverse_ttest(dat = dat, method_missing = 'none')
            overview_nomissing[[iter]] <- do.call(cbind, list(pvalues_nomissing[[iter]],
options_sample_size[[i_sample_size]], options_missing_prop[[i_missing]],
options_nIndicator[[i_indicators]], options_loadings[[i_loadings]]))
          }
          else{
            pvalues_withmissing[[iter]] <- multiverse_ttest(dat = dat)
            overview_withmissing[[iter]] <- do.call(cbind, list(pvalues_withmissing[[iter]],
options_sample_size[[i_sample_size]], options_missing_prop[[i_missing]],
options_nIndicator[[i_indicators]], options_loadings[[i_loadings]]))
          }
```

```
        iter <- iter + 1
      }
    }
  }
}
}
end <- proc.time()
end - start
results_nomissing <- do.call(rbind, overview_nomissing)
colnames(results_nomissing) <- c("mean p-value", "harmonic p-value", "sample_size",
"missing_prop", "nIndicator", "loadings")
results_withmissing <- do.call(rbind, overview_withmissing)
colnames(results_withmissing) <- c("mean p-value", "harmonic p-value", "sample_size",
"missing_prop", "nIndicator", "loadings")


# Check for unique values and combine datasets
length(unique(results_nomissing[,1]))
length(unique(results_withmissing[,2]))
results_combined <- do.call(rbind, list(results_nomissing, results_withmissing))


# Put the results in a table (power)
# Table of results_withmissing
results_withmissing <- as.data.frame(results_withmissing)
iter = 1
tab_results_withmissing <- matrix(NA, nrow = 16, ncol = 3)
names <- matrix(NA, nrow = 16, ncol = 1)
colnames(tab_results_withmissing) <- c("power of mean p-value", "power of harmonic mean
p-value", "number of observations")
for (i in 1:length(options_sample_size)){
  for (j in 1: length(options_missing_prop)){
    for (k in 1: length(options_nIndicator)){
      for (l in 1:length(options_loadings)){
```

```
        tab_results_withmissing[iter, 1] <-
sum(results_withmissing[results_withmissing$sample_size == options_sample_size[i] &
results_withmissing$missing_prop == options_missing_prop[j] &
results_withmissing$nIndicator == options_nIndicator[k] & results_withmissing$loadings ==
options_loadings[l], 1] <=
0.05)/length(results_withmissing[results_withmissing$sample_size ==
options_sample_size[i] & results_withmissing$missing_prop == options_missing_prop[j] &
results_withmissing$nIndicator == options_nIndicator[k] & results_withmissing$loadings ==
options_loadings[l], 1])
        tab_results_withmissing[iter, 2] <-
sum(results_withmissing[results_withmissing$sample_size == options_sample_size[i] &
results_withmissing$missing_prop == options_missing_prop[j] &
results_withmissing$nIndicator == options_nIndicator[k] & results_withmissing$loadings ==
options_loadings[l], 2] <=
0.05)/length(results_withmissing[results_withmissing$sample_size ==
options_sample_size[i] & results_withmissing$missing_prop == options_missing_prop[j] &
results_withmissing$nIndicator == options_nIndicator[k] & results_withmissing$loadings ==
options_loadings[l], 2])
        tab_results_withmissing[iter, 3] <-
length(results_withmissing[results_withmissing$sample_size == options_sample_size[i] &
results_withmissing$missing_prop == options_missing_prop[j] &
results_withmissing$nIndicator == options_nIndicator[k] & results_withmissing$loadings ==
options_loadings[l], 1])
        names[iter] <- c(paste0("sample_size = ", options_sample_size[i], " missing_prop = ",
options_missing_prop[j], " nIndicator = ", options_nIndicator[k], " loadings = ",
options_loadings[l]))
        iter = iter + 1
      }
    }
  }
}
rownames(tab_results_withmissing) <- names
kable(tab_results_withmissing)
```

```
# Table of results_combined
results_combined <- as.data.frame(results_combined)
iter = 1
tab_results_combined <- matrix(NA, nrow = 16, ncol = 3)
names <- matrix(NA, nrow = 16, ncol = 1)
colnames(tab_results_combined) <- c("power of mean p-value", "power of harmonic mean p-value", "number of observations")
for (i in 1:length(options_sample_size)){
  for (j in 1: length(options_missing_prop)){
    for (k in 1: length(options_nIndicator)){
      for (l in 1:length(options_loadings)){
        tab_results_combined[iter, 1] <- sum(results_combined[results_combined$sample_size == options_sample_size[i] & results_combined$missing_prop == options_missing_prop[j] & results_combined$nIndicator == options_nIndicator[k] & results_combined$loadings == options_loadings[l], 1] <= 0.05)/n
        tab_results_combined[iter, 2] <- sum(results_combined[results_combined$sample_size == options_sample_size[i] & results_combined$missing_prop == options_missing_prop[j] & results_combined$nIndicator == options_nIndicator[k] & results_combined$loadings == options_loadings[l], 2] <= 0.05)/n
        tab_results_combined[iter, 3] <- length(results_combined[results_combined$sample_size == options_sample_size[i] & results_combined$missing_prop == options_missing_prop[j] & results_combined$nIndicator == options_nIndicator[k] & results_combined$loadings == options_loadings[l], 1])
        names[iter] <- c(paste0("sample_size = ", options_sample_size[i], " missing_prop = ", options_missing_prop[j], " nIndicator = ", options_nIndicator[k], " loadings = ", options_loadings[l]))
        iter = iter + 1
      }
    }
  }
}
rownames(tab_results_combined) <- names
kable(tab_results_combined)
```

Appendix C – Multiverse_cor function

```
multiverse_cor <- function(nFactor = 2, nIndicator = 3, method_missing = c("listwise", "pairwise", "mi"), method_outliers = c("none", "IQR_remove", "z_remove"), dat = create_data()){
  result = list()
  if (sum(method_missing %in% c("listwise", "pairwise","mi") & "none" %in% method_missing ) >= 1){
    stop("Invalid combination")
  }
  if ("IQR_remove" %in% method_outliers){
    IQR_dat <- dat
    for (i in 1:(nIndicator * nFactor)) {
      IQR_dat[, i][IQR_dat[, i] > summary(IQR_dat[, i])[5] + 1.5 * IQR(IQR_dat[, i], na.rm = T) | IQR_dat[, i] < summary(IQR_dat[, i])[2] -1.5 * IQR(IQR_dat[, i], na.rm = T)] <- NA
    }
  }
  if("z_remove" %in% method_outliers){
    z_dat <- dat
    scaled_dat <- scale(z_dat)
    for (i in 1:(nIndicator * nFactor)){
      z_dat[, i][scaled_dat[, i] > 3 | scaled_dat[, i] < -3] <- NA
    }
  }
  if ("none" %in% method_missing){
    if (sum(is.na(dat)) != 0){
      warning("The data has missing values")
    }
  }
  if ("none" %in% method_missing & "none" %in% method_outliers){
    result[[1]] <- corr.test(dat[, 1:nIndicator], dat[, (nIndicator + 1):(nIndicator*nFactor)], adjust = "none")$p
  }
  if ("none" %in% method_missing & "IQR_remove" %in% method_outliers){
```

```
  result[[2]] <- corr.test(IQR_dat[, 1:nIndicator], IQR_dat[, (nIndicator +
1):(nIndicator*nFactor)], adjust = "none")$p
 }
 if ("none" %in% method_missing & "z_remove" %in% method_outliers){
  result[[3]] <- corr.test(z_dat[, 1:nIndicator], z_dat[, (nIndicator + 1):(nIndicator*nFactor)],
adjust = "none")$p
 }
 if ("listwise" %in% method_missing & "none" %in% method_outliers){
  complete_dat <- dat[complete.cases(dat),]
  result[[4]] <- corr.test(complete_dat[, 1:nIndicator], complete_dat[, (nIndicator +
1):(nIndicator*nFactor)], adjust = "none")$p
 }
 if ("listwise" %in% method_missing & "IQR_remove" %in% method_outliers){
  IQR_complete_dat <- IQR_dat[complete.cases(IQR_dat),]
  result[[5]] <- corr.test(IQR_complete_dat[, 1:nIndicator], IQR_complete_dat[, (nIndicator
+ 1):(nIndicator*nFactor)], adjust = "none")$p
 }
 if ("listwise" %in% method_missing & "z_remove" %in% method_outliers){
  z_complete_dat <- z_dat[complete.cases(z_dat),]
  result[[6]] <- corr.test(z_complete_dat[, 1:nIndicator], z_complete_dat[, (nIndicator +
1):(nIndicator*nFactor)], adjust = "none")$p
 }
 if ("pairwise" %in% method_missing & "none" %in% method_outliers){
  result[[7]] <- corr.test(dat[, 1:nIndicator], dat[, (nIndicator + 1):(nIndicator*nFactor)],
adjust = "none", use = "pairwise")$p
 }
 if ("pairwise" %in% method_missing & "IQR_remove" %in% method_outliers){
  result[[8]] <- corr.test(IQR_dat[, 1:nIndicator], IQR_dat[, (nIndicator +
1):(nIndicator*nFactor)], adjust = "none", use = "pairwise")$p
 }
 if ("pairwise" %in% method_missing & "z_remove" %in% method_outliers){
  result[[9]] <- corr.test(z_dat[, 1:nIndicator], z_dat[, (nIndicator + 1):(nIndicator*nFactor)],
adjust = "none", use = "pairwise")$p
 }
```

```
  if ("mi" %in% method_missing & "none" %in% method_outliers){
    capture.output(imputed <- mice(data = dat, seed = 2346982))
    df.imp <- complete(imputed)
    result[[10]] <- corr.test(df.imp[, 1:nIndicator], df.imp[, (nIndicator +
1):(nIndicator*nFactor)], adjust = "none")$p
  }
  if ("mi" %in% method_missing & "IQR_remove" %in% method_outliers){
    capture.output(IQR_imputed <- mice(data = IQR_dat, seed = 2346982))
    IQR_df.imp <- complete(IQR_imputed)
    result[[11]] <- corr.test(IQR_df.imp[, 1:nIndicator], IQR_df.imp[, (nIndicator +
1):(nIndicator*nFactor)], adjust = "none")$p
  }
  if("mi" %in% method_missing & "z_remove" %in% method_outliers){
    capture.output(z_imputed <- mice(data = z_dat, seed = 2346982))
    z_df.imp <- complete(z_imputed)
    result[[12]] <- corr.test(z_df.imp[, 1:nIndicator], z_df.imp[, (nIndicator +
1):(nIndicator*nFactor)], adjust = "none")$p
  }
  if (sum(method_missing %in% c("listwise", "pairwise","mi")) >= 1 & sum(is.na(dat)) ==
0){
    warning("The data has no missing values")
  }
  result_tot <- do.call(rbind, result)
  output <- matrix(NA, ncol = 2)
  output[,1] <- mean(result_tot, na.rm = T)
  output[,2] <- p.hmp(result_tot[! is.na(result_tot)],L=length(result_tot[! is.na(result_tot)]),
multilevel=FALSE)
  print(output)
}
```

Appendix D – Justification for assuming MCAR

In creating the multiverse for the current study, we focused on creating datasets that had missing data. However, given the nature of our multiverse we had to think about missing data in a different way. For example, missing at random (MAR) can be realistically assumed when one has missing data in their dataset, but this was deemed illogical in the setting of our generic multiverse analysis. Suppose that the indicator variables on the first factor represent different IQ tests which all have the same mean. In this case, it would not make sense for a person who scored high on one IQ test to have more missing data on (one of) the other tests/variables. This is also true if we look at the data in our multiverse from a different perspective. For example, a researcher could attempt to measure general knowledge (latent factor) by means of a trivia questionnaire. Instead of the variables being three different tests like in the IQ example, the variables could be different ways of measuring correct/incorrect answers to the trivia questionnaire. Having a high number of correct answers on the first coding key (i.e., the first variable) could not possibly lead to having more missing data on the other two variables, since it would still be the same test. Meanwhile, missing not at random (MNAR) and missing completely at random (MCAR) would still be possible in these examples. Here, MNAR could occur when people who have a lower IQ score or people who score poorly on the trivia questionnaire would skip questions as a result, thus leading to more missing data. It would also be possible to have missingness that is MCAR. For example, due to data entry mistakes by the researcher or due to the computer not saving an answer correctly. In the current study, we chose to solely focus on missingness that is MCAR. While MNAR is certainly possible, we believe it is less likely to be assumed by researchers since it is often difficult to determine whether the missingness is indeed MNAR.