Pegem Journal of Education and Instruction, Vol. 11, No. 3, 2021 (pp. 88-101)

RESEARCH ARTICLE WWW.PEGEGOG.NET

Beck Depression Inventory-II: A Study for Meta Analytical Reliability Generalization

Mehmet Taha ESER1*a, Gökhan AKSU2*b

[a]Adnan Menderes University, Education Faculty, Aydın/Turkey

[b]Adnan Menderes University, Education Faculty, Aydın/Turkey

Abstract

The main aim of achieving with the reliability generalization is to investigate the variability related to the reliability estimates and to try to characterize the sources of this variability. As part of the research, a reliability generalization study was carried out on the basis of Beck Depression Inventory-II to investigate potential factors contributing to the variability of the reliability of the measurement results and to examine the sources of the measurement error. Within the scope of the study, it was published in English between 2011-2019 and only 40 articles in the type of article were examined. The Kappa coefficient for the coding form was determined to be 0.93 and it was concluded that the measurement results performed for the coding form were valid and reliable. Jamovi and R programs were used in the research. When the test results regarding publication bias are evaluated in a holistic way, it is concluded that there is no publication bias related to the studies included in the research. It was thought that the heterogeneity observed by the researchers may indicate an amount of heterogeneity to be examined and moderator analyzes were performed. As a result of the moderator analysis, it was determined that any of the continuous and categorical moderator variables did not have an explanatory role regarding the variability between the reliability estimates of the inventory. In order to carry out qualified RG studies in the future, it is recommended that researchers report their reliability estimates regarding the measurement results of their studies.

Keywords: Meta-Analysis, Reliability Generalization, Beck Depression Inventory-II, VC Model, Cronbach Alpha.

1. Introduction

The concept of depression is characterized by worthlessness, decreased interest in daily activities, self-worthiness, depressive mood, decreased concentration or focus, lack of motivation and thoughts of suicide or death (American Psychiatric Association, 2013). Researchers working in the field of psychology, psychiatry, and education include the concept of depression in their work. Usually, in the field of psychology and psychiatry the concept of depression is included in the studies targeting the development of interventions that will best address mental health problems, whereas in the field of education, the objective is to determine the role of depression in educational processes, and measurement tools for depression are used in this context. The review of the literature showed that the Beck Depression Inventory (BDI) is one of the most popular measurement tools that help measure the severity of the depression, along with the Center for Epidemiological Studies Depression Scale and the Hamilton Depression Rating Scale (Bentz & Hall, 2008; McDowell, 2006)

Beck Depression Inventory II (BDI-II)

BDI, developed by Beck et al. in 1961, has undergone two major revisions (Beck, Rial, & Rickels, 1974). The first revision was carried out in 1978 (BDI-IA) and the second in 1996 (BDI-II). BDI-II is the revised version of BDI-I according to DSM IV criteria. In BDI-II, weight loss, physical imaging, loss of workability and somatic complaints were removed from the inventory. Instead, agitation, concentration difficulty, worthlessness thoughts and energy loss were added. BDI-II measures somatic, emotional, cognitive and motivational symptoms of depression. It is available in different forms, including the computer form and the card form. The inventory is based on data from clinical observations and is a 21-item measurement tool that is not based on a specific theory. 4-points scale of 0-3 range was used. The scores that can be achieved from the inventory vary between 0-63. There are over 10 studies for the adaptation of different cultures. Depression is not diagnosed using the results obtained from BDI, the severity of depression symptoms is determined objectively (Hisli, 1989; Savaşır & Şahin, 1997).

Beck Depression Inventory-II and Reliability

It is known by the researchers that the reliability estimates of the measurements vary according to the sample characteristics,

Corresponding Author e-mail: m.taha.eser@adu.edu.tr

https://orcid.org/orcid.org/ 0000-0001-7031-1953

How to cite this article: Mehmet Taha Eser MT, Aksu G (2021). Beck Depression Inventory-II: A Study for Meta Analytical Reliability Generalization. Pegem Journal of Education and Instruction, Vol. 11, No. 3, 2021, 88-101

Source of support: Nil

Conflict of interest: None.

DOI: 10.14527/pegegog.2021.00

Submission : 15.09.2020 **Revision:** 25.04.2021

Acceptence: 28.04.2021 **Publication:** 01.07.2021

working conditions and score distributions. Researchers are recommended to report the reliability estimates of their studies results. Such transparent reporting practices provide critical information necessary for the researchers to consider scoring reliability while interpreting study results, and for other researchers to make informed decisions about the applicability of data set and study results. When researchers appropriately report psychometric data about the administration of a measurement tool, score reliability may be analyzed multiple times to understand how the measurement error may change under fluctuating working conditions. Such an approach requires the quantitative integration of reliability coefficients that are suitable for meta-analytical methods such as Reliability Generalization (RG).

Regarding BDI II; Cronbach alpha reliability coefficient indicating internal consistency is α= .92; test-retest reliability at one-week interval is r = .93; Convergent validity given by Beck Anxiety Inventory is r =.56; and discriminant validity given by the Sociotropy-Autonomy Scale is r =-. 10 (Beck, Steer, & Brown, 1996; Steer & Clark, 1997).

Even though BDI has undergone more reliability testing than BDI-II, both inventories are considered to be highly reliable (Dozois, Dobson, & Ahnberg, 1998). The original manual of BDI-II reported high internal consistency with a coefficient of 0.93 for university students and a coefficient of 0.92 for patients with psychiatric outpatients (Beck et. al., 1996).

More recently, Dozois & Covin (2004) reviewed 13 works that reported reliability data of BDI-II since 1996 and reported an average Cronbach alpha coefficient of 0.91. Although less information is available on the test-retest reliability of BDI-II, in the original manual 1-week test-retest reliability coefficient was reported to be 0.93 for 26 psychiatric clinics (Beck et al. 1996). However, as pointed by Dozois & Covin (2004), it is difficult for test-retest reliability to both reliably measure depression and to be interpreted as a measure that detect changes in treatment-related depression. For example, one group of researchers suggested that BDI may not be reliable for a longer period of time in non-clinical samples after finding that the BDI scores of a non-clinical sample decreased by 40% in 2 months (Ahava, Iannone, Grebstein, & Schirling, 1998). Such a significant drop in BDI scores in non-clinical samples clearly threatens the tool's ability to reliably detect the changes in depression arising solely from the treatment.

Considering the BDI-II versions with different formats and different number of items and the extensive use of the inventory in different types of samples, it becomes necessary to examine whether the psychometric properties of the inventory, especially the reliability of the measurements, can be generalized. Since 1998, more than 100 RG works, in which a wide variety of meta-analytical and statistical methods were used, have been published (Holland, 2015). A researcher who is willing to use RG in his/her study should make several methodological decisions, including the selection of statistical models for the synthesis of reliability coefficient and moderator analysis, as well as the transformation and weight of the coefficients within these models.

Reliability Generalization

Traditionally, two types of statistical models are used within the scope of meta-analysis studies, namely the fixed effects model (Hedges & Olkin, 1985) and the random effects model (Hedges & Vevea, 1998; Hunter & Schmidt, 2004). Classical fixed effects models are based on 2 assumptions: all coefficients of the study are estimated for the same sample and any deviation from the parameter results in sampling error (Bonett, 2010; Hedges, 1992). In general, it is recommended to use fixed effects models when it is desired to generalize the results to similar works within the scope of meta-analysis. Fixed effects methods have been found to perform poorly under many meta-analysis-specific conditions, and these models are generally not recommended for routine use (Bonett, 2008; Rodriguez & Maeda, 2006; Schmidt, Oh, & Hayes, 2009). Random effects models are based on the assumption that more than one sample parameters are involved, and each work included in the meta-analysis represents an example of a hypothetical sample of the past or future works. Therefore, each coefficient is considered as an estimate of its own sample's parameter that can vary from work to work (Bonett, 2008; Rodriguez & Maeda, 2006). Random effects models tend to establish very wide confidence intervals compared to fixed effects model because the additional error is explained by the variance between the works (Meca, López-López, & López-Pina, 2013). It is not recommended to use random effects models within the scope of meta-analysis studies involving reliability coefficient due to the bias that may occur in parameter estimates, failure to achieve interpretable estimates about parameter variance and violation of assumptions about sampling method (Rodriguez & Maeda, 2006). On the other hand, Laird & Mosteller (1990) suggested the use of the Varying Coefficient Meta-Analytical Method (VC), which is a method involved in the meta-analysis of the reliability coefficient, and Bonett used the VC model for the meta-analysis of Cronbach Alfa (2010). VC retains the advantages of both the fixed and random effects model and is seen as an alternative to these two models. As a fixed effects model, results obtained using VC can be generalized only to the works similar to those included in meta-analysis. However, instead of assuming that all of the alpha estimates are equal to a single constant parameter, each work is assumed to estimate its sample's reliability coefficient, similar to the fixed effects approach. The size of the error components under the VC model is moderate and the confidence intervals generated by the VC model are between those estimated by fixed effects or random effects models (Bonett, 2010; Sánchez-Meca et. al., 2013). VC model provides more accurate confidence intervals

than fixed effects or random effects model; it provides excellent results in parameter estimation in small samples; and it can be used in a much wider range of problems than traditional models (Bonett, 2010).

RG is a meta-analytical method for synthesizing the reliability coefficients of the measurements of different works (Caruso, 2000; Vacha-Haase, 1998). The main objective of RG is to investigate the variability of the reliability estimates and to try to characterize the sources of this variability (Vacha-Haase, Henson & Caruso, 2002). RG is an extension of the validity generalization (Schmidt & Hunter, 1977). The works in which RG is used allow researchers to have an idea about estimating the expected measurement error, as well as the effect size, power and statistical significance of future works (Henson & Thomson, 2002; Nimon, Ziantek, & Henson, 2012). Researchers are suggested to carry out the RG study by using the reliability values of a few, carefully selected and high-quality works in order to obtain accurate estimates about the expected reliability and to determine the potential effects of the moderator variables (Bonett, 2010). In this context, the results of RG studies encourage meta-analytical thinking, which serves to establish a historical and contextual framework to better evaluate a single work's result (Henson, 2006; Thompson, 2002). Using meta-analysis in a study by including statistical information of previous works in an existing work helps to create radical changes in behavioral sciences as well as establishing an integrative and cumulative scientific process (Henson, 2006; Bonett, 2010).

The point that RG is based on is that according to the review of the reliability coefficients of the measurements obtained from different sample sizes, test lengths, test forms, test versions and test administration conditions, the measurement results are reliable, not the tests. RG helps researchers and practitioners to better understand the measurement tools they use in their studies. Researchers can make informed decisions by knowing the most suitable measurement tools for certain types of samples, and the sub-scales whose measurement reliability is below the acceptable level. In this way, researchers and testers get more detailed information about the measuring tool that should be used under certain conditions (Barnes, Harp, & Jung, 2002; Kieffer, 1999; Kieffer & Reese, 2002). RG studies can also generate measurement results with high effect size and power. It is notable that only about 10% of the researchers reported the reliability that they calculated for the measurement tool they used in their research (Barnes, Harp, & Jung, 2002; Beretvas, Meyers, & Leite, 2002). RG studies encourage researchers to report the reliability of their measurements (Vacha-Haase & Thompson, 2011). Otherwise, the reliability of the measurements of the published works cannot accurately represent the performed works and "file drawer problem" arises. General limitations of RG studies can be listed as: reliability values are not reported within the scope of the screened works; Cronbach's alpha values are reported for the reliability; the researcher (s) conducting the RG study employs reliability transformation calculations (Henson & Thompson, 2002).

According to the literature Although BDI has been developed for clinical use, the literature review showed that BDI and its versions have been used on different samples over the years (Barrera & Garrison-Jones, 1988; Hatzenbuehler, Parpal, & Mattews, 1983). The figures related to the reliability of the measurements reported in the works were observed to vary according to BDI versions used in different types of researches. Without using the meta-analytical RG approach, it is very difficult to understand why the reliability of the measurements related to BDI-II varies. In this context, as it is one of the most popular measurement tools that objectively determine the severity of depression symptoms, investigating potential factors contributing to the variability of the measurement's reliability and to examine the sources of the measurement error is thought to be important in informing the researchers about RG and guiding them on how to perform an RG study, considering that no meta-analytical RG study was available in Turkish literature. Regarding the purpose and importance of the research, the works, in which the relevant criteria available in the BDI II coding form have been reported, were included in the sample. RG method was used in this study, and Bonett's VC model was preferred as the model.

Method

Research Design

This research can be reviewed within the scope of RG studies, which is a meta-analytical method that examines the sources of error variance, considering multiple works that use a particular measuring tool or measuring tool group that measure the same structure (Vacha-Haase, Henson & Caruso, 2002). In RG studies, reliability of the works, characteristics of the sample and the characteristics of the measurement tool are used as predictors, and the relationships between the dependent variables and predictors are analyzed to explain the variability of the reliability coefficients (Mason, Allam, & Brannick, 2007; Vacha-Haase, 1998).

Participants

Publications in the Web of Science database have been screened. In determining the keywords, the criteria of the works to be included in the research were considered; if the number of keywords is too high, the quality of the work and the reliability would decrease, and if it is too rigid, we would fail to generalize the results because of including too few works (Lam & Kennedy, 2005). The screening was preformed using the following keywords: "Beck Depression Inventory II and Reliability"; "Beck Depression Inventory-II and Reliability",

"Beck Depression Inventory II Reliability", "Beck Depression Inventory-II Reliability", 'BDI II Reliability' and 'BDI-II Reliability'. 40 publications, which fulfill the following criteria, were included in the study: Cronbach Alfa reliability coefficient should be reported to minimize the amount of error without the need to perform coefficient transformation (Thompson & Vacha-Haase, 2000); the sample size, men and women percentage, average score and standard deviation should be reported; publication year should be between 2011-2019 to ensure that the dataset consist of current works; and the publication language should be English and should not be included in the Turkish literature to ensure that the works are international.

Considering the psychometrics theory, the reliability of the measurements is expected to be affected by variables such as the group's characteristics and standard deviation of group measurements (Nunnally, 1978). In the criteria selection process, the literature providing information about potential variables recommended to be used as criteria in RG studies were reviewed (Henson & Thompson, 2002; Vacha-Haase & Thompson, 2011). The first screening was carried out on 10.12.2019; the last screening on 21.12.2019. The criteria used to examine the possible relationships between the estimates of measurements' reliability of the works included in the study and the characteristics of the works were coded as follows and used as moderator variables:

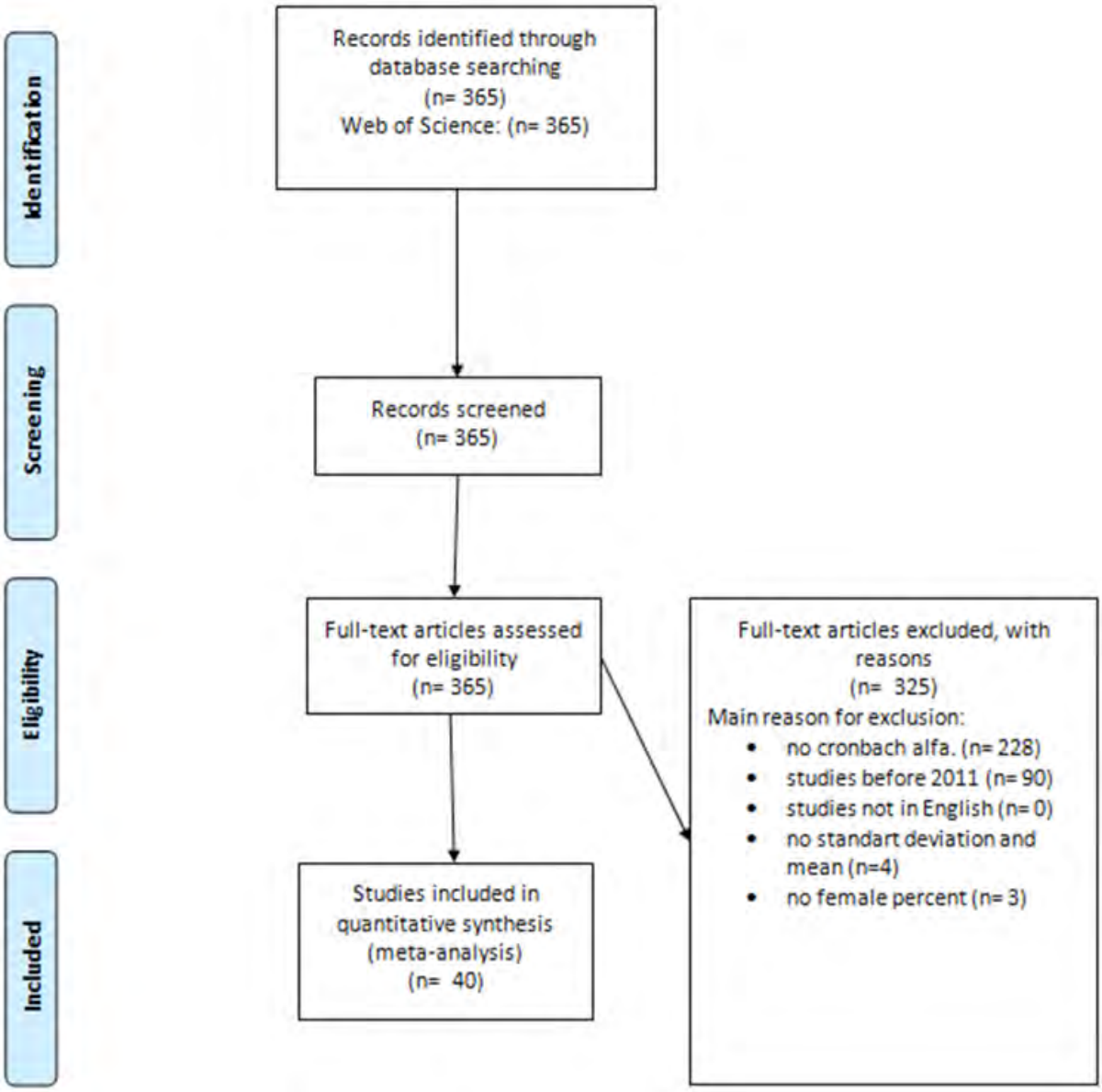

Fig. 1: Flow chart of the data collection process

Table 1.: The criteria and the works included in the meta-analysis.

Author	*Year*	*N*	*Number of Items*	*Cronbach*	*Female %*	*Mean*	*SD*	*Language*	*Sample Type*
Toledano and Valdez (2018)	1	446	21	0.90	83.00	13.88	9.83	1	0
Jesus Sanz1 (2013)	0	712	21	0.90	80.02	19.98	10.96	1	0
Jesus Sanz2 (2013)	0	569	21	0.87	90.39	9.61	7.76	1	2
Jesus Sanz3 (2013)	0	727	21	0.89	91.25	8.75	7.34	1	1
Campos-Gonçalves (2011)	0	538	21	0.90	60.00	8.90	7.90	1	1
Gorenstein1 (2011)	0	3410	21	0.88	71.00	10.90	8.20	1	1
Gorenstein2 (2011)	0	1417	21	0.89	60.00	11.70	9.30	1	1
Gorenstein3 (2011)	0	301	21	0.89	61.00	10.40	10.10	1	1
Gorenstein4 (2011)	0	182	21	0.93	56.00	9.90	10.70	1	1
Dolle et al. (2012)	0	88	21	0.94	58.00	10.50	8.90	1	0
Sashidharan (2012)	0	278	21	0.91	75.00	9.40	3.60	0	1
Roberts1 (2012)	0	115	21	0.90	82.00	5.10	5.90	1	1
Roberts2 (2012)	0	37	21	0.96	60.00	5.10	5.90	1	0
Whisman (2013)	0	7369	21	0.90	65.00	9.30	8.10	0	1
Corbiere (2011)	0	206	21	0.84	53.00	17.20	11.50	1	0
Hayden (2012)	0	83	21	0.89	71.00	13.40	9.10	0	0
Bunevicius et al. (2012)	0	522	21	0.85	28.00	11.00	8.20	1	0
Kirsch-Darrow (2011)	0	161	21	0.89	31.00	9.50	7.20	0	0
Lopez (2013)	0	345	21	0.93	0.00	23.00	12.20	0	0
Tully (2011)	0	226	21	0.85	17.00	8.60	6.20	0	0
Turner (2012)	0	72	21	0.94	47.00	13.40	12.90	0	0
Williams1 (2012)	0	136	21	0.90	33.00	6.50	5.20	0	1
Williams2 (2012)	0	93	21	0.90	35.00	14.70	7.40	0	0
Brouwer et al. (2012)	0	1530	21	0.90	62.00	20.10	10.80	1	0
Gonzalez et al. (2017)	1	391	21	0.92	39.60	9.31	7.84	1	1
Eun-Ho Lee et al. (2017)	1	1072	21	0.89	51.00	9.63	7.19	1	1
Dadfar and Kalibatseva (2016)	1	52	21	0.85	73.10	11.30	7.55	1	0
Dahem1 (2016)	1	250	21	0.79	100.00	27.17	10.78	1	1
Dahem2 (2016)	1	250	21	0.73	0.00	28.74	10.63	1	1
Garcia-Batista et al. (2018)	1	1040	21	0.89	54.10	16.91	11.62	1	2
Oliveira et al.1 (2012)	0	182	21	0.93	51.00	9.87	10.71	1	1
Oliveira et al.2 (2012)	0	80	21	0.92	0.00	7.88	9.12	1	1
Oliveira et al.3 (2012)	0	102	21	0.93	100.00	11.43	11.62	1	1
Henndy Ginting et al. (2013)	0	720	21	0.90	28.80	11.60	8.10	1	0
Mahmoudi et al. (2019)	1	138	21	0.86	54.50	8.20	5.82	1	2
Yu-Mai Song et al. (2012)	0	1967	21	0.88	55.10	11.33	7.97	1	1
González et al.1 (2015)	1	391	21	0.92	60.40	9.31	7.84	1	1
González et al.2 (2015)	1	205	21	0.87	57.10	9.82	7.70	1	1
Khine La Win1 (2019)	1	40	21	0.91	5.00	19.90	13.60	1	0
Khine La Win2 (2019)	1	186	21	0.93	5.40	14.40	9.20	1	1

Note: Works with numbers at the end of the author's name indicate the measurements for different sample sizes and (or) types within the same work.

- Publication Year (2011-2014: 0; 2015-2019: 1).
- Percentage of women in the sample (50% and below: 0, above 50%: 1)
- Average of the measurements (Continuous variable)
- Standard deviation of the measurements (Continuous variable)
- The language of the measuring tool (English: 0; Other: 1).
- Sample type (Patient: 0; Healthy: 1; Patient and healthy: 2).

Information about the criteria and the works included in the meta-analysis is shown in Table 1.

Validity and Reliability of the Measurement Related to the Coding Form

As a result of the literature screening, 40 works fulfilling the criteria were included in the study. The works included in the research were coded through the form created by the researchers. The coding form includes the following information: the name of the work, the author (s) of the work, the year of the work; sample size, number of items, reliability coefficient, percentage of women in the study, average of sample's measurements; standard deviation of the sample's measurements, language of the measurement tool and the sample type. In order to ensure the content validity of the coding form, 3 experts, who are lecturers in Educational Sciences, were given detailed information about the steps of the research and expert opinion was received. In line with the expert opinion, a code indicating sample type was added to the coding form. Beside the scope validity of the coding form, inter-coder reliability was also checked. For this purpose, randomly selected ten studies were coded by two independent coders and Cohen's Kappa Coefficient (κ) was calculated as the cohesion criterion between coders. The Kappa coefficient corrects the part of the rapport between the coders based on chance and gives information about the real cohesion rate (Sim & Wright, 2005). As a result of the calculation, κ was found to be 0.93. This result can be interpreted as "almost perfect" (Altman, 1999). Considering these results, it is concluded that the measurements of the coding form are valid and reliable.

Regarding the objectives of the research, the study group was formed by the students in the age group of 15 who were registered in formal education and who participated in PISA 2015 test organized by OECD. A total of 540000 students from 72 countries have participated in the test, and 5895 of them were from Turkey. Regarding the execution of PISA 2015 in Turkey, the population of the students in the age group of 15 consisted of 1324089 students; whereas accessible population was 925366 students. In PISA research, school sample was determined according to stratified random sampling method (MoNE, 2016).

Data Analysis

Jamovi 1.1.9.0 and R-3.6.2 programs were used in the analyzes carried out within the scope of the study. Jamovi (2018) is a new, free software on R language, based on popular R packages for the analysis, which are performed through drop-down menus; whereas R is an old (1993), free and popular software (Eser, Yurtçu, & Aksu, 2020). Metaphor package developed by metaphor (Viechtbauer, 2010) was used in both software and rma function was used in this context. Many of the statistical models used within the scope of meta-analytical RG studies are based on the assumption that the effect sizes show a normal distribution (Rodriguez and Maeda, 2006). Cronbach alpha statistics taking values in the range of 0-1 violates this assumption. The reliability coefficients are normalized by using Bonett Transformation (ln (1-α)) to meet the assumption considering the number of studies and sample sizes (Bonett, 2010).

In qualitative studies, study group should be preferred instead of sample since such studies are conducted with few individuals or units. The individuals or units forming the study group should be introduced with all relevant characteristics. Information regarding the context of the study group should also be explained here.

Limitations of the Study

There are a few limitations related to this study. Only the works written in English, published between 2011-2019 and paper-type works were included in the study. In addition, the analyzes carried out within the scope of the study are limited to the variables included in the coding.

Findings

Descriptive Statistics of the Works

Regarding the descriptive statistics related to the works included in the study, in which research criteria were reported; 8 of the works were published in 2011, 16 in 2012, 4 in 2013, 2 in 2015, 3 in 2016, 2 in 2017, 2 in 2018 and 3 in 2019. The total sample size of the works included in the study was 26,629. The English version of the inventory was used in 9 of the works, while the remaining 31 were in other languages. The sample of 16 works was comprised of patients, 21 of healthy individuals and the remaining 3 was comprised of both patients and healthy individuals.

Tables and figures can be used to display the results of the analyses. Findings section should deal only with presenting the results and should not include the discussion of the findings. Sub-headings in line with sub-goals of the study can be used. Sub-headings should be flush left, in italics and with each word capitalized.

Findings about Normality, Publication Bias, Model Fit and Confidence Intervals

In order to perform a meta-analytical RG with raw Cronbach Alpha values, which are the reliability coefficients of the works included in the study, these coefficients should exhibit a normal distribution. Since Cronbach Alpha coefficients are doubly bounded variable, the distributions of these coefficients are not normal, therefore they are not suitable for direct modeling by meta-regression, which is based on the normal distribution assumption. Therefore, the RG study should be performed using the coefficients obtained by performing Bonett transformation (Bonett, 2002; Rodriguez & Maeda, 2006). Although it can be theoretically supported that Cronbach Alpha values do not show normal distribution, the normality of distribution of alpha coefficients was also tested in the study. Shapiro-Francia, Anderson-Darling and Shapiro-Wilk tests were used to test the normality of the alpha coefficients' distribution. The results of these tests are given in Table 2.

Table 2: Normality test results.

Test	*Value*	*p*
Shapiro-Francia	0.9523	0.0850
Anderson-Darling	0.6661	0.0759
Shapiro-Wilk	0.9635	0.2203

Regarding the statistics related to the normal distribution in Table 2, p values were observed to be above .05 significance level for all statistical tests. Based on this result, it was concluded that the reliability coefficients of the works included in the study did not show a normal distribution. After determining that the reliability coefficients do not show a normal distribution, the existence of publication bias for the works included in the study was checked. The results of the funnel plot indicating the possibility of publication bias are given in Figure 2.

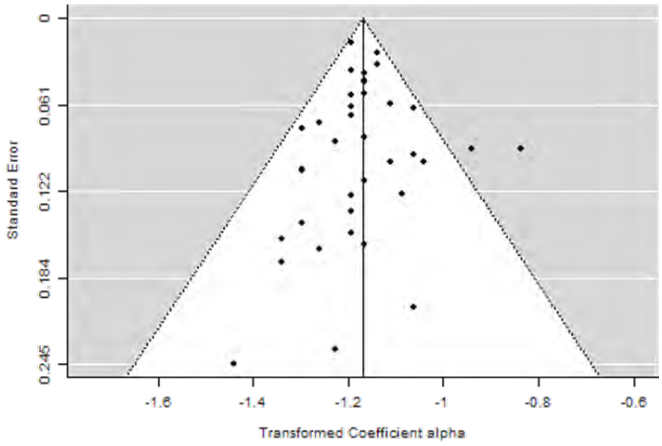

Fig. 2: Funnel plot of Cronbach alpha coefficients

Regarding the funnel plot in Figure 2, the works included in the study were observed to be grouped in the middle section and the black dots located on the right and left of the vertical line, which symbolize the combined effect size, show an almost symmetrical scattering. Apart from the funnel chart, the existence of publication bias was also examined by interpreting the result of the regression test for the asymmetry of the funnel chart and Kendall's Tau coefficient. As a result of the holistic evaluation of the visual illustration of the funnel graph, the result of regression test for the asymmetry of the funnel chart ($z = -0.551$; $p = 0.58 > 0.05$) and Kendall's Tau coefficient ($T = -0.122$; $p = 0.268 > 0.05$), it was concluded that there was no publication bias regarding the studies included in the research. Mullen, Muellerleile and Bryant (2001) stated that the results of the meta-analysis studies can be resistant to future researches if the value obtained by using the formula $N/(5k + 10)$ is greater than 1. The value obtained as a result of the calculation made using the relevant formula, was found to be greater than 1 ($26629 / (40 * 5 + 10) = 126.804$). Based on this result, it can be said that the publication bias of the meta-analytical RG study is very low.

After testing the normality assumption and analyzing publication bias, the results of the analysis about the heterogeneity of the works included in the study were tested. Considering that Bonett's VC model is basically a random effects model (Holland, 2015); this study does not include all works in which BDI-II was employed as the sampling method; and the random effects model is a more realistic representation of the real world (Field, 2003b), Bonett's VC were used in the analysis. The restricted maximum-likelihood method was used to test the heterogeneity of variance, because it shows low-level bias and is an estimation method suitable for the use of the works with small and large samples together (Langan et al., 2019). Although restricted maximum-likelihood method was preferred as the estimation method because of its advantages, model statistics comparing restricted maximum likelihood method and maximum likelihood method are given in Table 3.

Table 3.: Model fit statistics and information criteria.

Criteria	*log-likelihood*	*AIC*	*BIC*
Maximum-Likelihood	38.020	-74.039	-72.350
Restricted Maximum-Likelihood	36.061	-70.121	-68.458

Regarding Table 3, model fit statistics of the restricted maximum likelihood method were observed to be smaller than these of maximum likelihood method. Considering that model fit is provided for smaller values of log-likelihood, AIC and BIC (Fabozzi, Focardi, Rachev, & Arshanapalli, 2014), it was concluded that the model fit was in favor of the restricted maximum-likelihood method. After the analysis of the model fit, an analysis was performed for calculating the average reliability and the lower and upper limits of reliability in 95%

Table 4: Basic Output for VC Model (k = 40)

		Average Reliability (Estimate)	*se*	*Z*	*p*	*CI Lower Bound*	*CI Upper Bound*
αBonett		-1.17	0.0008	-132	< .001	-1.189	-1.154
Cronbach	α	0.898	0.0009	976	< .001	0.896	0.900

Note: αBonett represents the values obtained through Bonetti transformation, whereas Cronbach α represents the raw reliability values.

confidence interval. Considering the difficulty of interpreting the values obtained from Bonett transformation, Table 4, in which the results of the analysis are shown, includes both the values obtained from Bonett transformation and Cronbach Alpha reliability coefficients.

As a result of the analysis performed based on the VC Model, the average reliability in terms of Cronbach Alfa was found to be 0.898 with an error of 9.21e-4, whereas the lower and upper limits were found to be 0.896 and 0.900 in 95% confidence interval. The review of the average reliability and the lower and upper limits of the confidence interval according to Taber (2017) showed that they can be considered as reliable because they are in the range of 0.84-0.90.

The forest plot, created according to the Cochrane Review Forest Plot Template, showing the distribution of the reliability of the works included in the study is given in Figure 3.

The red squares in Figure 3 indicate the effect size (reliability) of the studies, and the lines next to the squares indicate the upper and lower limits of their effect size in 95% confidence interval. In addition, the weight percentage given on the left side of the forest plot shows the effect share of each work on the meta-analysis result numerically. The work with the highest weight, that is, with the largest red square, is the work of Whisman (2013) within the scope of both fixed effects and random effects model. Therefore, it can be said that work of Whisman (2013) has the largest effect size. The work with the smallest red square is the work of Khine La Win1 (2012) and it can be said that this work is the one with the smallest effect size. Regarding the forest plot in Figure 3, the work with the largest confidence interval was observed to be Whisman's work (2013), whereas the one with the narrowest confidence interval belongs to Oliviera et al.3

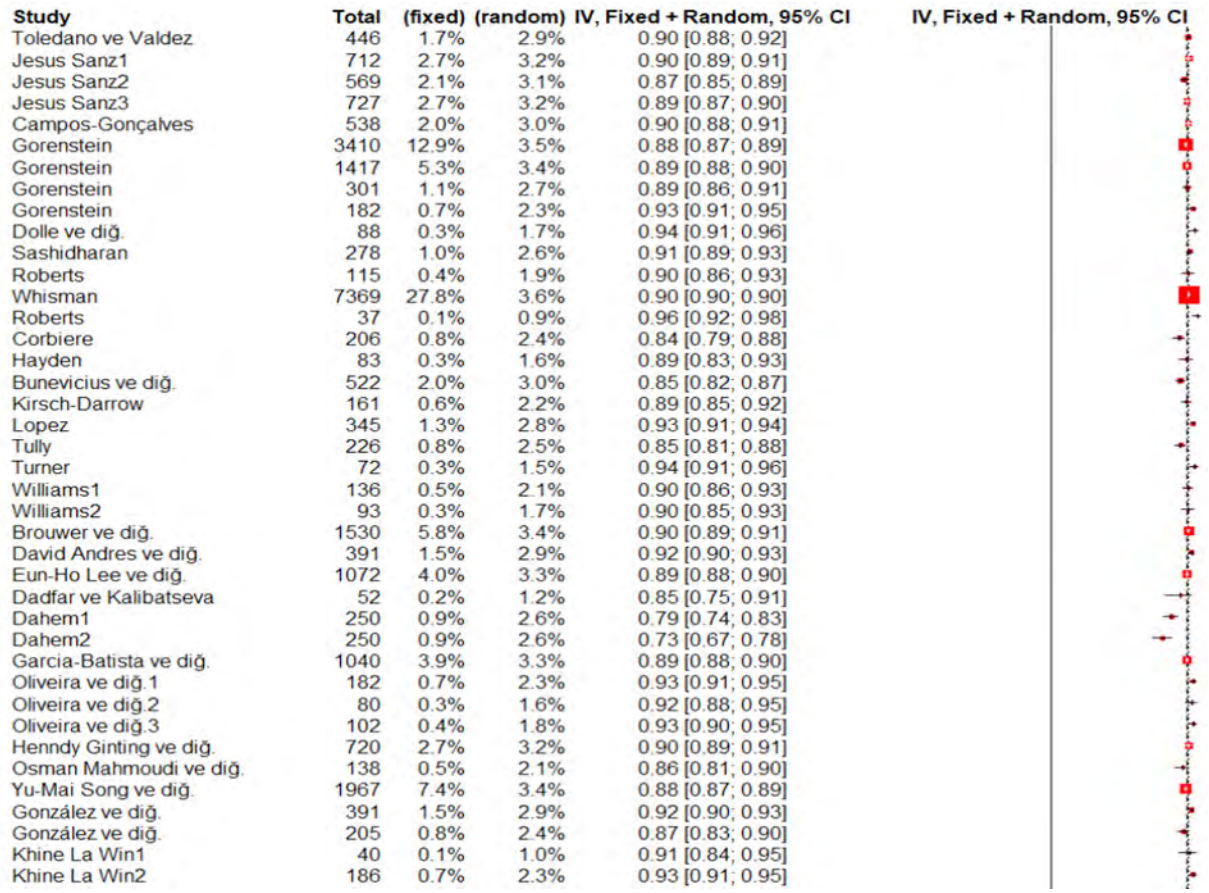

Fig. 3: Forest plot regarding the reliability of the works.

(2012). Regarding the analysis results about the weights of the works, Whisman's (2013) work was observed to have largest weight (27.8%), whereas the works of Roberts (2012) and Khine La Win1 (2019) have the smallest weight (0.1%). The weights of the other works seems to be similar.

Regarding the analysis results about the heterogeneity, it was observed that (Q (df = 39) = 47.6373, p-val = 0.1615) and I2 = 18.13146. It is necessary to make an explanation at this point regarding the Q statistic, which is generally used to indicate the amount of heterogeneity in the literature, and the I2 statistic, which is generally used in the literature by classifying as low-medium-high.

The Q value of a heterogeneity test is a function of (1) the amount of observed heterogeneity, (2) the accuracy of individual works, and (3) the number of works in the analysis. The Q value may be large if the estimated heterogeneity is insignificant. On the contrary, it may be small is the estimated heterogeneity is important. Therefore, the Q-value should never be used instead of the amount of heterogeneity. Regarding I2, which is the other statistics on heterogeneity, the idea of using this statistic for classifying heterogeneity as low, medium or high is meaningless for two reasons. Firstly, I2 is not an absolute distribution index, it is a ratio. It does not provide information about the distribution. Secondly, a heterogeneity that can be considered high in one context can be considered as low in another (Borenstein, 2019).

Considering the low number of works covered in the research and the purpose and importance of the research, even though the Q value was not statistically significant, and the I2 statistic has not received a relatively large value, the narrowness of the lower and upper limit range of the alpha value in Table 4 ([0.896-0,900]) let the researchers think that the observed heterogeneity may indicate an amount of heterogeneity that should be tested and moderator analyzes were carried out to guide future Meta-Analytical RG researches.

Moderator Analysis Findings Regarding Whether the Reliability Coefficient Differs According to The Publication Year, Average and Standard Deviation of Sample Scores

Within the scope of the moderator analysis, firstly, continuous moderator variables were included in the meta-regression and the analyzes were carried out. Each moderator variable was added to the meta-regression separately. Table 5 shows the findings related to meta-regressions performed with moderator variables.

Table 5 shows the meta-regression results of continuous moderators. Regarding Table 5, the regression coefficients of the continuous moderator variables included in the meta-regressions were not statistically significant, that is, none of the variables can be said to play a role in the change of reliability. Regarding R2 values, indicating the explained variance of the variables, the explained variance for all variables was observed to be zero. R2 value(s) equal to zero for the findings of the moderator analysis is quite common and the literature review shows many studies having R2 value (s) equal to zero (Meca, López-Pina, López-López, Marín-Martínez, Rosa-Alcázar, & Gomez-Conesa, 2011; Rubio-Aparicio, Núñez-Núñez, Sánchez-Meca, López-Pina, Marín-Martínez, & López-López, 2020; Vicent, Rubio-Aparicio, Sánchez-Meca, & Gonzálvez, 2019; Vassar & Bradley, 2012). Q statistics giving information about heterogeneity were observed to be statistically insignificant and I2 rates were very low.

Moderator Analysis Findings Regarding Whether the Reliability Coefficient Differs According to The Language of the Inventory, The Sample Type and The Publication Year

To address the second sub-problem of the moderator analyzes, categorical moderator variables were included in the meta-regression and the analyzes were carried out. Each moderator variable was added to the meta-regression separately. Table 6

Table 5: Results of the Simple Meta-Regression Analyses Assuming VC Model for Transformed Alpha Coefficients of Continuous Moderator Variables.

Moderator Variable	*k*	*bj*	*se*	*z*	*p*	*QE*	*I2*	*R2*
Female %	40	0.0010	0.0005	0.0736	0.9414	QE(38) = 47.6356; p=0.1359	8,39	0
Average	40	0.0041	0.0021	1.9392	0.0525	QE(38) = 43.8769; p=0.2364)	9,90	0
Standard Deviation	40	-0.0015	0.0061	-0.2519	0.8011	QE(38) = 47.5739; p=0.1373)	8,52	0

Note: k = number of studies; bj = unstandardized regression coefficient; z= significance test of the regression coefficient; p = p value of the significance test. QE = statistic to test the model misspecification. In order to facilitate the interpretation, the regression coefficients were transformed back to the metric of the original coefficients, I2=heterogeneity index, R2 = contribution of the moderator variable to the explained variance

Table 6: Results of the ANOVA Applied on Transformed Alpha Coefficients for the Version Language, Population Type and Publication Year

	k_j	α_+	*95% CI* $[\alpha_l; \alpha_u]$	I^2	*ANOVA* Results
Version Language					
Others	9	0.8644	[0.8126; 0.9163]	0.00	F(1. 38) = 3.3957. p=0.0732 R2= 0.00 QE(38) = 43.7296. p= 0.2412
English	31	0.8997	[0.8898; 0.9096]		
Population Type					
Patient	16	0.8969	[0.8779; 0.9158]	9.69	F(2. 37) = 0.3694. p=0.6937 R2 = 0.00 QE(37) = 46.7025. p=0.1318
Healthy	21	0.8905	[0.8686; 0.9123]		
Patient and Healthy	3	0.8733	[0.8354; 0.9113]		
Publication Year					
2011-2014	28	0.9004	[0.8894; 0.9113]	8.71	F(1.38) = 0.6025. p= 0.4424 R2 = 0.00 QE(38) = 46.8497. p=0.1537
2015-2019	12	0.8717	[0.8341; 0.9092]		

Note. K_j= number of studies; $\alpha_{+=}$ mean alpha coefficient; F= Knapp–Hartung's statistic for testing the significance of the moderator variable; Q_w= the statistic for testing the model misspecification; I^2= heterogeneity index; R^2 = contribution of the moderator variable to the explained variance.

shows the findings related to meta-regressions performed with moderator variables.

Table 6 shows the meta-regression results of categoric moderators. Regarding Table 6, F-test results of the analysis carried out with categoric moderator variables were observed to be statistically insignificant. Regarding R2 values indicating the explained variance, the explained variance was observed to be zero. Q statistics giving information about heterogeneity were observed to be statistically insignificant and I2 rates were very low.

The review of Table 5 and Table 6 together shows that none of the continuous and categorical moderator variables have an explanatory role for the variability of the reliability estimates of the inventory. At the same time, Q statistics, which provide information about the presence of heterogeneity, were not statistically significant for continuous and categorical moderator variables; I2 statistics were observed to have very small values. In Meta Analytical RG studies, it is recommended to establish an estimation model including the variables that play a role in the change of reliability by considering the moderator analysis performed by including continuous and categorical variables (Rubio-Aparicio et al., 2020). Creating a model plays a very important role in explaining the variability of reliability coefficients. Multiple Meta-Regression Analysis is used to define the subset of the most relevant characteristics of the works (moderators) to explain the variability of alpha coefficients. Considering that none of the continuous and categorical moderator variable has an explanatory role in the variability of alpha coefficients, multiple meta-regression analysis was not performed.

Discussion, Conclusion & Implementation

Within the scope of this study, a meta-analytical RG was performed to estimate the reliability of the scores obtained using BDI-II and to determine the characteristics that are statistically related to the variability of the reliability coefficients. We also estimated the reliability induction rate and compared the characteristics of the works reporting and inducing reliability. The RG carried out within the scope of the research is based on a total of 40 studies.

The tendency to over-generalize the reliability coefficients of a measuring tool's previous measurements has been named as "reliability induction" by Vacha-Haase et al. (2000). Reliability induction causes many researchers to omit the calculation and reporting of the reliability coefficients related to the measurements and the researches are based on previously published data of different data sets (Willkinson and The APA Task Force on Statistical Inference, 1999). Only 137 (37%) out of 365 papers selected to be included in the meta-analytical RG study reported reliability estimates obtained for the samples. For the remaining 228 (63%) works, 53 (23%) of them only stated that BDI-II is a reliable measurement tool; whereas 175 (77%) reported reliability estimates found in other works, using reliability induction.

In order to avoid mixing the reliability coefficients calculated by different reliability concepts (test-retest, Omega etc.), this study focused on Cronbach Alpha values obtained from 40 papers. The average alpha value obtained from the reliability estimates of these 40 works was 0.898 with an error

of 0.009 and this value is in an acceptable reliability range (Nunnally & Bernstein, 1994). The lower and upper limits of the confidence interval were found to be 0.896 and 0.900. The Cronbach Alpha value reported in the original paper involving the development of BDI-II was 0.92 (Beck, Steer & Brown, 1999), which is higher than the average reliability and upper limit obtained in the study. This is thought to be related to the use of scores obtained from a single sample, which is used to determine BDI-II factor structure in the original paper.

The review the analysis results related to heterogeneity within the scope of the research (Q(df = 39) = 47.6373, p-val = 0.1615; I2= 18.13146) made the authors of this study think that the values should be further examined and moderator analyzes were performed considering that the reliability estimates may not explain sampling error alone. According to the results of the moderator analysis, none of the continuous (female participant percentage, average, standard deviation) and categorical (language, sample type, publication year) moderator variables play an explanatory role in the variability of BDI-II's reliability estimates. Considering the effect of standard deviation and average value on the alpha coefficients in psychometrics theory (Nunnally & Bernstein, 1994) and the effects of the sampling characteristics on the reliability, the results of the research may seem to be contrary to the theory of psychometrics, but the literature review revealed that there are other studies in which the moderator variables used in meta-analytical RG were observed to have no role in explaining the change in reliability (Lopez-Pina et al., 2009; Meca et al., 2011; Vassar and Bradley, 2012; Rubio-Aparicio et al., 2020; Vicent et al., 2019). Based on this, it was concluded that the average reliability and lower-upper limit values can be generalized for BDI-II. But it should be kept in mind that reliability is a characteristic of the scores, not a characteristic of the measurement tool.

Although it has been 24 years since its development, BDI-II continues to be used frequently in related studies. Although there are other measurement tools frequently used for the same purpose in the literature such as BDI-II and other measurement tools were developed for the same purpose, it is thought that the widespread and extensive use of BDI-II will continue.

In addition to the RG studies available in the literature, it is recommended to conduct new studies and write reference books. Reliability will be better understood theoretically as more RG studies are conducted and published; The psychometric information obtained as a result of the increase in sample diversity and awareness will be cumulatively used and more evidence will be collected on features that are thought to affect the reliability of the measurement results.

Reporting of the researchers' data reliability estimates is considered to be important. In the research, only Cronbach Alpha values were used for reliability. In future studies, the coefficients obtained by different reliability methods can be combined in a single study using transformation formulas. This study did not examine the validity of BDI-II scores in various works or samples. Therefore, validity should not be directly evaluated according to the results of the study.

The data collected for RG studies is comprised of the reliability values included in the works on the relevant subject. In order to carry out high-quality RG studies in the future, researchers are recommended to report the reliability estimates of the measurements in their studies.

BDI-II consists of two subscales, cognitive-affective subscale and a somatic-performance subscale. Reliability reports on subscales were found to be very few in number by the authors of this study and they were not covered by the study. For future RG studies, it is recommended to include the reliability values related to the sub-scales that constitute the measurement tools.

References

Ahava, G. W., Iannone C., Grebstein, L., & Schirling J. (1998). Is the Beck Depression Inventory reliable over time? An evaluation of multiple test-retest reliability in a nonclinical college student sample. *Journal of Personality Assessment*, *70*, 222-231.

Altman, D. G. (1999). *Practical Statistics for Medical Research*. Chapman; Hall/CRC Press.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: Author.

Barnes, L. L. B., Harp, D., & Jung, W. S. (2002). Reliability Generalization of Scores on the Spielberger State-Trait Anxiety Inventory. *Educational and Psychological Measurement*, *62*, 603-618.

Barrera, M., & Garrison-Jones, C. V. (1988). Properties of the Beck Depression Inventory as a screening instrument for adolescent depression. *Journal of Abnormal Child Psychology*, *16* (3), 263–273.

Beck, A. T., Steer, R. A. & Brown, G. K. (1996). Manual for the Beck Depression Inventory-II. San Antonio, TX: Psychological Corporation.

Beck, A. T., Rial, W. Y., & Rickels, K. (1974). Short form of depression inventory: cross-validation. *Psychol Rep., 34* (3), 1184-1186.

Bentz, B. G., & Hall, J. R. (2008). Assessment of depression in a geriatric inpatient cohort: A comparison of the BDI and GDS. *International Journal of Clinical and Health Psychology*, *8* (1), 93–104.

Bonett, D. G. (2010). Varying coefficient meta-analytic methods for alpha reliability. *Psychological Methods*,*15*, 368–385.

Bonett, D. G. (2008). Meta-analytic interval estimation for bivariate correlations. *Psychological Methods*, *13* (3), 173-189.

Borenstein, M. (2019). *Common mistakes in meta-analysis and how to avoid them*. Biostat, Inc, Englewood, NJ.

Brouwer, D, Meijer, R. R., & Zevalkink, J. (2013). On the Factor Structure of the Beck Depression Inventory-II: G Is the Key. *Psychol Assess., 25* (1), 136-145.

Bunevicius, A., Staniute, M., Brozaitiene, J., & Bunevicius, R. (2012). Diagnostic accuracy of self-rating scales for screening of depression in coronary artery disease patients. *Journal of Psychosomatic Research*, *72* (1), 22-25.

Campos, R. C., & Gonçalves, B. (2011). The Portuguese version of the Beck Depression Inventory-II (BDI-II): preliminary psychometric data with two nonclinical samples. *European J Psychol Assess. 27* (4), 258-264.

Caruso, J. C. (2000). Reliability generalization of the NEO personality scales. *Educational and Psychological Measurement*, *60* (2), 236-254.

Corbière, M., Bonneville-Roussy, A., Franche, R. L., Coutu, M. F., Choinière, M., Durand, M. J., & Boulanger, A. (2011). Further validation of the BDI-II among people with chronic pain originating from musculoskeletal disorders. *The Clinical Journal of Pain, 27* (1), 62-69.

Dadfar, M., & Kalibatseva, Z. (2016). Psychometric Properties of the Persian Version of the Short Beck Depression Inventory with Iranian Psychiatric Outpatients. *Scientifica,* 1-6.

Dahem, F. (2016). Psychometric Properties of the Beck Scale for Depression (Beck Depression Inventory BDI-II) - A Study on a Sample of Students in the State of Kuwait Universities. *Journal of Education and Practice, 7* (17), 87-99.

Dolle, K, Schulte-Körne, G, O'Leary, A. M., Von Hofacker, N, Izat, Y, & Allgaier, A. K. (2012). The Beck Depression Inventory-II in adolescent mental health patients: Cut-off scores for detecting depression and rating severity. *Psychiatry Res. 200* (2), 843-848.

Dozois, D. J. A., & Dobson, K. S. & Ahnberg, J. L. (1998). A psychometric evaluation of the Beck Depression Inventory-II. *Psychological Assessment, 10*(2), 83–89.

Dozois, D. J. A., & Covin, R. (2004). The Beck Depression Inventory-II (BDI-II), Beck Hopelessness Scale (BHS), and Beck Scale for Suicide Ideation (BSS). In M. J. Hilsenroth & D. L. Segal (Eds.), *Comprehensive handbook of psychological assessment, Vol. 2. Personality assessment* (pp. 50–69). New York: John Wiley & Sons Inc.

Eser, M. T., Yurtçu, M., & Aksu, G. (2020). *R programlama dili ve Jamovi ile meta analiz uygulamaları*. Ankara: Pegem Akademi.

Fabozzi, F.J., Focardi, S., Rachev, S.T., & Arshanapalli, B. (2014). *The basics of financial econometrics: Tools, concepts, and asset management applications*. Wiley.

Field, A. P. (2003b). The problems in using fixed effects models of meta-analysis on real-world data. Understanding Statistics, *2*, 77 – 96.

García-Batista, Z. E., Guerra-Peña, K., & Cano-Vindel, A., Herrera-Martínez, S. X., & Medrano, L. A. (2018). Validity and reliability of the Beck Depression Inventory (BDI-II) in general and hospital population of Dominican Republic. *PLoS One. 13* (6), 1-12.

Ginting, H., Näring, G., Williams, V. V., Srisayekti, W., & Becker, E. (2013). Validating the Beck Depression Inventory-II in Indonesia's general population and coronary heart disease patients. *International Journal of Clinical and Health Psychology, 13*, 235–242.

Gomes-Oliveira, M. H., Gorenstein, C., Lotufo Neto, F., Andrade, L. H., & Wang, Y. P. (2012). Validation of the Brazilian Portuguese version of the Beck Depression Inventory-II in a community sample. *Braz J Psychiatry. 34* (4), 389-394.

González, D. A., Reséndiz, A., & Reyes-Lagunes, I. (2015). Adaptation of the BDI-II in Mexico. *Salud mental, 38* (4), 237–244.

Gorenstein, C, Wang Y. P., Argimon, I. L, & Werlang, B. S. G. (2011*). Manual do Inventário de Depressão de Beck - BDI-II*. São Paulo: Casa do Psicólogo.

Hatzenbuehler, L. C., Parpal, M., & Matthews, L. (1983). Classifying college students as depressed or nondepressed using the Beck Depression Inventory: An empirical analysis. *Journal of Consulting and Clinical Psychology*, *51* (3), 360–366.

Hayden, M. J., Brown, W. A., & Brennan, L., & Brien, P. E. (2012). Validity of the Beck Depression Inventory as a Screening Tool for a Clinical Mood Disorder in Bariatric Surgery Candidates. *Obesity Surgery, 22* (11), 1666-1675.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3* (4), 486–504.

Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, *7* (2), 246-255.

Henson, R. K. (2006). Effect-size measures and meta-analytic thinking in counseling psychology research. *The Counseling Psychologist*, *34* (5), 601-629.

Henson, R. K., & Thompson, B. (2002). Characterizing measurement error in scores across studies: Some recommendations for conducting "reliability generalization" studies. *Measurement and Evaluation in Counseling and Development*, *35*, 113-127.

Hisli, N. (1989) Beck Depresyon Envanteri'nin üniversite öğrencileri için geçerliği güvenirliği. *Psikoloji Dergisi*, *23*, 3-13.

Holland, D. F. (2015). *Reliability Generalization: A Systematic Review And Evaluation Of Meta-Analytic Methodology And Reporting Practice* (Doctoral dissertation, North Texas University, Texas, USA). Retrieved November 13, 2020, from https://digital.library.unt.edu/ark:/67531/metadc822810/

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis*. Thousand Oaks, CA: Sage.

Kieffer, K. M. (1999). Why Generalizability Theory is Essential and Classical Test Theory is Often Inadequate. Thompson, B. (Ed.), *Advances in Social Science Methodology* (pp. 1-11), Stamford, Connecticut: JAI.

Kieffer, K. M. ve Reese, R. J. (2002). A Reliability Generalization Study of the Geriatric Depression Scale. *Educational and Psychological Measurement*, *62*, 969- 994.

Kirsch-Darrow, L., Marsiske, M., Okun, M. S., Bauer, R., & Bowers, D. A. (2011). Apathy and depression: separate factors in Parkinson's disease. *The Journal of the International Neuropsychological Society, 17* (6), 1058-1066.

Laird, N. M., & Mosteller, F. (1990). Some Statistical Methods for Combining Experimental Results. *International Journal of Technology Assessment in Health Care, 6* (1), 5-30.

Lam, R. W., & Kennedy, S. H. (2005). Using meta analysis to evaluate evidence: Practical tips and traps. Canadian Journal of Psychiatry, *50* (3), 167-174.

Langan, D., Higgins, J., Jackson, D., Bowden, J., Veroniki, A., Kontopantelis, E., Viechtbauer, W., & Simmonds, M. (2019). A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research Synthesis Methods*, *10* (1), 83-98.

Lee, E. H., Lee, S. J., Hwang, S. T., Hong, S. H., & Kim, J. H. (2017). Reliability and Validity of the Beck Depression Inventory-II among Korean Adolescents. *Psychiatry Investigation, 14* (1), 30-36.

Lopez, M. N., Pierce, R. S., Gardner, R. D., & Hanson, R. W. (2013). Standardized Beck Depression Inventory-II scores for male veterans coping with chronic pain. *Psychological Services, 10* (2), 257–263.

Mahmoudi, O., Paydar, M., Amini, M. R., Mohammadi, F., & Darvishi, M. (2019). Beck Depression Inventory: Establishing

the Reliability and Validity of the Kurdish Version Among Earthquake Survivors of Kermanshah, Iran. *International Journal of Health and Life Sciences, 5* (1), 1-5.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory.* New York, NY: McGraw-Hill, Inc.

Rubio-Aparicio, M., Núñez-Núñez, M. R., Meca, J. S., López-Pina, A. J., Marín-Martínez, F., & López-López, A. J. (2020) The Padua Inventory–Washington State University Revision of Obsessions and Compulsions: A Reliability Generalization Meta-Analysis, *Journal of Personality Assessment*, *102* (1), 113-123.

Mason, C., Allam, R., & Brannick, M.T. (2007). How to meta-analyze coefficient-of-stability estimates: Some recommendations based on Monte Carlo studies. *Educational and Psychological Measurement*, *67* (5), 765-783.

McDowell, I. (2006). *Measuring health: A guide to rating scales and questionnaires*. New York: Oxford University.

Mullen, B., Muellerleile, P. ve Bryant, B. (2001). Cumulative meta-analysis: A consideration of ındicators of sufficiency and stability. Personality and Social Psychology Bulletin, *27* (11), 1450-1462.

Nimon, K., Zientek, L. R., & Henson, R. K. (2012). The assumption of a reliable instrument and other pitfalls to avoid when considering the reliability of data. *Frontiers in Quantitative Psychology and Measurement*, *3*, 1-13.

Nunnally, J. C. (1978). *Psychometric theory.* New York: McGraw-Hill.

Odriozola-González, P., & Ruiz, F. J. (2016). The role of psychological inflexibility in Beck's cognitive model of depression in a sample of undergraduates. *Anales de Psicología, 32* (2), 441-447.

Roberts, G., Roberts, S., Tranter, R., Whitaker, R., Bedson, E., Tranter, S., Prys, D., Owen, Heledd & Sylvestre, Y. (2012). Enhancing rigour in the validation of patient reported outcome measures (PROMs): bridging linguistic and psychometric testing. *Health and Quality of Life Outcomes, 10* (64), 1-6.

Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. Psychological Methods, *11* (3), 306-322.

Meca, J. S., López-López, J. A., & López-Pina, J. A. (2013). Some recommended statistical analytic practices when reliability generalization studies are conducted. *British Journal of Mathematical and Statistical Psychology*, *66* (3), 402-425.

Meca, J. S., López-Pina, J. A., López-López, J., Marín-Martínez, F., Rosa-Alcázar, A. I., & Gomez-Conesa, A. I. (2011). The Maudsley Obsessive-Compulsive Inventory: A reliability generalization meta-analysis. *International Journal of Clinical and Health Psychology, 11* (3), 473-493.

Sanz, J. (2013). 50 Years of The Beck Depression Inventory: Recommendatıons for Usıng the Spanısh Adaptation of the BDI-II in Clinical Practıce. *Papeles del Psicólogo, 34* (3), 161-168.

Sashidharan, T., Pawlow, L. A., & Pettibone, J. C. (2012). An examination of racial bias in the Beck Depression Inventory-II. *Cultur Divers Ethnic Minor Psychol., 18* (2), 203-209.

Savaşır, I., & Şahin N. H. (1997) *Bilişsel Davranışçı Terapilerde Değerlendirme: Sık Kullanılan Ölçekler.* Ankara: Türk Psikologlar Derneği Yayınları.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, *62* (5), 529-540.

Schmidt, F. L., Oh, I.S., & Hayes, T. L. (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology*, *62*, 97-128.

Sim, J. & Wright, C. C. (2005). The Kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, *85* (3), 257-268.

Steer, R. A, & Clark, D.A. (1997). Psychometric characteristics of the Beck Depression Inventory-II with college students. *Measurement and Evaluation in Counseling and Development*, *30*, 128–136.

Taber, K. S. (2017). The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education*, *48* (6), 1273–1296.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*, 25-32.

Thompson, B. & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement*, *60* (2), 174– 95.

Toledano-Toledano, F., & Contreras-Valdez, J. A. (2018). Validity and reliability of the Beck Depression Inventory II (BDI-II) in family caregivers of children with chronic diseases. *PLoS ONE, 13* (11), 1-13.

Tully, P. J., Winefield, H. R., Baker, R. A., Turnbull, D. A., & De Jonge, P. (2011). Confirmatory factor analysis of the Beck Depression Inventory-II and the association with cardiac morbidity and mortality after coronary revascularization. *Journal of Health Psychology, 16* (4), 584–595.

Turner, A., Hambridge, J., White, J., Carter, G., Clover, K., Nelson, L., & Hackett, M. (2012). Depression screening in stroke: A comparison of alternative measures with the structured diagnostic interview for the diagnostic and statistical manual of mental disorders, fourth edition (major depressive episode) as criterion standard. *Stroke, 43* (4), 1000-1005.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*, 6- 20.

Vacha-Haase, T., Henson, R. K., & Caruso, J. C. (2002). Reliability generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement*, *62* (4), 562-569.

Vacha-Haase, T., & Thompson, B. (2011). Score reliability: A retrospective look back at 12 years of reliability generalization studies. *Measurement and Evaluation in Counseling and Development*, *44* (3), 159-168.

Vassar, M., & Bradley, G. (2012). A reliability generalization meta-analysis of coefficient alpha for the Reynolds Adolescent Depression Scale. *Clinical Child Psychology and Psychiatry*, *17* (4), 519-527.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36* (3), 1-48.

Vicent, M., Rubio-Aparicio, M., Sánchez-Meca, J., & Gonzálvez, C. A. (2019). Reliability generalization meta-analysis of the child and adolescent perfectionism scale. J Affect Disord. *245*, 533-544.

Whisman, M. A., Judd, C. M., Whiteford, N. T., & Gelhorn, H. L. (2013). Measurement Invariance of the Beck Depression Inventory-Second Edition (BDI-II) across gender, race, and ethnicity in college students. *Assessment, 20* (4), 419-428.

Wilkinson, L. & Task Force on Statistical Inference, American Psychological Association, Science Directorate. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54* (8), 594–604.

Williams, J. R., Hirsch, E. S., Anderson, K., Bush, A. L., Goldstein, S. R., Grill, S., Lehmann, S., Little, J. T., Margolis, R. L., Palanci, J., Pontone, G., Weiss, H., Rabins, P., & Marsh, L. (2012). A comparison of nine scales to detect depression in Parkinson disease: which scale to use?. *Neurology, 78* (13), 998–1006.

Win, K. L., Kawakami, N., & Htet Doe, G. (2019). Factor structure and diagnostic efficiency of the Myanmar version BDI-II among substance users. *Annals of general psychiatry, 18* (12), 1-7.