How is depression valued?

Papageorgiou, Aikaterini

Chapter 6

Capturing the value of depression: Towards a scoring algorithm for a depression-specific preference-based measure

Katerina Papageorgiou
Paul FM Krabbe
Karin M Vermeulen
Maya J Schroevers
Anne M Stiggelbout
Erik Buskens
Adelita V Ranchor

Submitted

ABSTRACT

Background: The aim of this study was to derive a scoring algorithm for the McSad depression-specific classification system for future use in cost-utility analyses of depression interventions.

Methods: Thirty profiles describing depression states were generated based on the McSad and valued using the time-trade-off task. Regression modelling was used to develop a scoring algorithm to predict values for all states possibly generated by the McSad. The validity of the model was assessed through its goodness-of-fit (R^2) and ensuing coefficients.

Results: The sample consisted of 1268 participants, representative of the general Dutch population. Each of the thirty McSad states was valued on average by 169 participants. Regression coefficients were used to assign weights to each level of each domain, on which the scoring algorithm was based. However, as indicated by the rather low R^2 (.011), and non-significant regression coefficients, model performance was not satisfactory, i.e., did not explain variability in our data.

Conclusions: We report on preliminary results to develop a scoring algorithm for a depression-specific preference-based measure and discuss the relevant challenges.

INTRODUCTION

Cost-utility analyses are increasingly being used to inform health policy. In cost-utility analyses, preference-based measures are commonly used to capture the effectiveness of health care interventions (Drummond, Sculpher, Torrance, O'Brien, & Stoddart, 2005; Gold, Siegel, Russel, & et al., 1996). Preference-based measures typically consist of a classification system with multiple health-related quality of life (HRQoL) domains meant to describe the health state; and of a scoring algorithm used to assign a single 0 – 1 value to each state possibly generated by the classification system ("value set"). This value represents the "preference" for this state (with usually 0 equivalent to "dead" and 1 equivalent to "full health"). Preference-based measures such as the EQ-5D (Rabin & de Charro, 2001) and the SF-6D (Brazier & Fitzpatrick, 2002) are mostly used as their classification systems include generic HRQoL domains, such as mobility or self-care, relevant to the wide majority of health conditions and are therefore recommended as they allow for the comparison of different types of interventions. Using these measures, people report their health state, by choosing the level that best describes their functioning in the HRQoL domains. Then the scoring algorithms are used to indirectly assign a 0-1 value to this unique state. These algorithms are typically developed based on studies in which participants are asked to directly value a limited number of states generated by the classification system, and statistical techniques are then applied to estimate predictive models. Commonly, preference-based measures with accompanying value sets are used to sidestep direct valuation which is both complicated and time and consuming.

However, the suitability of generic preference-based measures to capture changes in specific health conditions has been questioned, and this is the reason for the development of condition-specific preference-based measures (Brazier et al., 2008; Grimison, Simes, Hudson, & Stockler, 2009; Lamers, Groot, & Buijt, 2007; Petrillo & Cairns, 2011; Revicki, Leidy, Brennan-Diemer, Sorensen, & Togias, 1998; Sundaram, Smith, Revicki, Elswick, & Miller, 2009). Their suitability has especially been debated in the case of mental health conditions, as the focus until now appears to have been on somatic dimensions. The McSad (Bennett, Torrance, Boyle, Guscott, & Moran, 2000) is the only internationally available classification system that comprises of depression-specific HRQoL domains, such as Emotion, Self-appraisal and Cognition. It has demonstrated its validity and has been found to perform better than a generic classification system to reflect depression (Papageorgiou, Vermeulen, Schroevers, Buskens, & Ranchor, 2013). An accompanying value set would enhance the application of the McSad in cost-utility analysis to evaluate depression interventions, which currently lags behind (Barrett, Byford, & Knapp, 2005). However, such a value set is currently not available.

The aim of the current study was to develop a scoring algorithm intended to generate a value set for the McSad depression-specific classification system. For this purpose, we used data on the general Dutch population's valuations of a number of hypothetical health states based on the McSad classification system. We anticipated that this study would yield a valuation algorithm for future studies evaluating depression interventions.

METHODS

The McSad depression-specific classification system

The McSad classification system comprises of six HRQoL domains specific to depression (Emotion, Self-appraisal, Cognition, Physiology, Behavior, and Role function), and there are four levels representing severity of dysfunctioning (1: no – 4: severe) for each domain. Thus it is possible to generate 4096 (4^6) unique descriptive profiles of depression states. For example, profile 232322 describes mild dysfunctioning in Emotion, Cognition, Behavior, and Role function and moderate dysfunctioning in Self-appraisal and Physiology. The McSad has demonstrated its validity in attaining assessments of depression states (Bennett, Torrance, Boyle, & Guscott, 2000). In this study we used the validated Dutch version of the McSad (Papageorgiou et al., 2013).

Selection of health states for direct valuation

Currently, there is no identified optimal method to select sample sizes and health states for direct valuations when developing scoring algorithms. When selecting a subset of states from the McSad for direct valuation, our main concerns were the following two. Firstly, to ensure, taking the sample size into account that for each state there would be a sufficient number of valuations. There is currently no agreement among researchers concerning number of valuations required for each state, so we employed a minimum of 80 valuations per states. Secondly, to select states that would reflect the different combinations of the four levels of dysfunctioning in the six domains. The orthogonal procedure in SPSS was therefore used, as usually performed in similar type of studies. This resulted in the generation of twenty-five depression states. Given the targeted sample size, we expected that it was possible to include more states while keeping the number of observations per state satisfactory. Therefore, we added five more states which we chose with the purpose to keep the level of dysfunctioning across the six domains more consistent (for example mild to moderate dysfunctioning in all six domains, rather than mild dysfunctioning in one domain and severe in another).

Valuation survey

Participants of the valuation study had to be over 18 years of age, to be able to understand Dutch and sign an informed consent. They were recruited by an international company maintaining a panel for marketing and academic research (Survey Sampling International). We aimed for a representative sample of the general Dutch population and for this purpose a stratified sampling strategy was employed with strata for gender, age, educational background and residence. From the 2378 invited to the study, 1840 (80.7%) met the inclusion criteria and 1268 (53.4%) completed the survey. The survey was administered online.

Data collection was performed in the period between December 2012 and January 2013. Previously the study protocol had been refined following an in depth pilot study among ten participants (Papageorgiou et al., 2014), and tested in a quantitative pilot study among two hundred members of the general population. A waiver was provided by the Medical Ethical Committee of the University Medical Center Groningen after reviewing of the protocol (M12.119685).

The thirty depression states were presented to participants as vignettes. Each participant valued four vignettes, selected at random and presented in random order. This randomization resulted in the unequal number of participants per state. Participants were asked to imagine living in the health state described in the vignette. Then a time-trade-off (TTO) method was used to elicit valuations for the thirty vignettes. TTO is the most common method eliciting valuations (Green, Brazier, & Deverill, 2000a), previously used for valuations of depression (Green, Brazier, & Deverill, 2000b; Koenig, Guenther, Angermeyer, & Roick, 2009). In pilot testing, TTO has been found appropriate for the valuation of depression states generated from the McSad (Papageorgiou et al., 2014). In the current study, participants were asked to choose between two options. The first option was to live in the - adverse - health state as described in the vignette for another 10 years and then die. The second option was to fully recover from this health state, but to live for less than 10 years. The method requires that the participant reports the number of years less (0-10) for which Choice A and Choice B seem equally desirable. Based on this, the value attached to this health states is calculated as: $1 - (x_{max}/10)$, where x_{max} is the maximum years the participant was willing to trade to live free of depression. The value can thus range from 0 to 1, with lower scores representing more negative valuations. Participants were first trained in using the TTO task by means of a vignette describing a health state related to asthma. A ping-pong titration method was used in this training state. To facilitate the task, an interactive horizontal scale representing the life years was used. As pilot testing indicated that the ping-pong titration method was not necessary after the first valuation, we decided to omit it in the main valuation task, so that participants could directly report the maximum number of years they would be willing to trade off in order to cure from depression. Participants had to complete each task to continue with the next one. An example of the TTO task as presented to participants can be found in Figure 1.

Imagine that you experience symptoms of depression.

- Feel more down than usual and don't enjoy things as usual.
- Don't feel very good about yourself these days and often see the down-side of everything.
- Have some trouble concentrating and remembering these days, and it seems harder to make decisions.
- Sleep is a little troublesome these days. Don't have quite the normal get up and go, and have less of an appetite.
- Things are more of a chore these days and at time feel sluggish or agitated.
- Able to function okay at work, home, school or with friends but often don't enjoy what you are doing, or feel more withdrawn lately.

Imagine that you could choose between two options

- **Option A:** You will live as described for 10 more years and then you will die.
- **Option B:** Depression will be cured and you will live in healthy mental condition. However, you will live for less than 10 years.

What would you choose? Would you trade any of your life years to cure depression?
What is the maximum number of years you would trade?

**Please drag the cursor in the scale to give your answer*

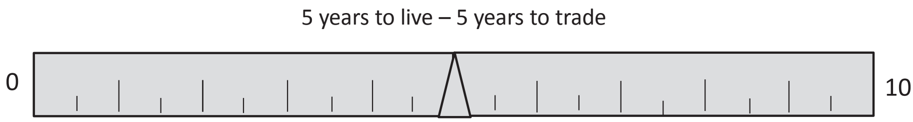

Figure 1 *Example of a TTO task as presented to participants*

Model development

There is currently no standardized prediction method available to model valuation data to arrive at an algorithm (Brazier, Ratcliffe, Salomon, & Tsuchiya, 2007; Dolan, 2002). We employed a simple additive model using ordinary least squares (OLS) regression models based on the individual level data. In this model, we used dummy variables to represent the specific level for each of the six domains as predictors and the value attached to the health state as the outcome variable. This means that there were 18 dummy variables (6 domains x 3 levels) used as predictors in the model and that a regression coefficient was produced for each. In order to predict a value for particular states generated by the McSad, the following formula was used: Predicted value = constant + (W1*D1 + W2*D2 + W3*D3 + W4*D4 + W5*D5 + W6*D6). In this model D1-6 represent the binary dummy variables recording the specific level in each of the six McSad domains that applies to this specific McSad state, whereas W1-W6 is the weight for this dummy variable, as indicated by the regression coefficients.

Model validation

In order to validate the estimated predictive model our first criterion was goodness-of-fit (produced R^2). Our second criterion was the statistical significance and the logical consistency (ordering and magnitude) of the produced regression coefficients. If, for example, coefficients produced for the levels in one domain were not significant, this implies that this domain is not a significant predictor of the outcome variable, i.e. valuations. Therefore, it is not contributing meaningfully to a scoring algorithm meant to predict valuations. Alternatively, if coefficients in one domain indicate that a level reflecting more severe dysfunctioning is related to less reduction in valuation outcome than a level reflecting less severe dysfunctioning, then the model appears logically inconsistent. When the aforementioned criteria are met, a third criterion was set, that is the predictive ability of the model. Predictive ability can be considered acceptable when the model produces small absolute differences between the TTO values and the predicted values for the thirty states, as generated by the scoring algorithm. All the analyses were performed using SPSS, version 16 (SPSS Inc., Chicago IL).

RESULTS

Data collection

Among the 2278 members of the general public who were invited to partake, 1268 (55.6%) participated and completed the survey. The final sample was representative of the general Dutch population with respect to age (mean (sd): 46.66 (17.41)), gender (51.1% female), residence and educational background. The majority of participants were married or in partnership (64.1%) and had children (58.8%). Average values attached to the thirty McSad states are presented in Table 1**,** ordered from the highest to the lowest values**.** Each state was valued on average by 169 participants (range: 155 – 183). For each of the 30 states, valuations ranged from 0 to 1. Average (SD) state values ranged from .58 (.31) to .72 (.33).

Model development & validation

The regression analyses included 18 dummy variables reflecting description of the McSad depression state (6 domains x 3 levels) used as predictors; and the values assigned to each state as the outcome variable. The adjusted R^2 of the simple linear regression model was rather low at .011 meaning that the model performs poorly in explaining the observed variability. Out of the 18 predictor dummy variables representing levels for the six domains, only four produced statistically significant coefficients, meaning that the majority of the levels in the McSad domains were not significant predictors of the value assigned (Table 2).

Table 1 *TTO-based and predicted mean values for all the 30 directly valued McSad states (ordered from the highest to the lowest)*

State	N	Mean	SD	Median	Predicted Value
111111	169	.72	.33	.85	.71
221112	169	.71	.29	.80	.70
411211	169	.69	.26	.70	.65
141113	176	.69	.28	.75	.67
114121	169	.68	.30	.80	.68
222222	162	.68	.30	.80	.66
322121	171	.67	.30	.70	.64
212413	174	.67	.29	.70	.68
141422	166	.66	.30	.70	.65
112242	160	.66	.31	.70	.66
211131	167	.65	.28	.70	.69
132311	155	.64	.30	.70	.68
223112	181	.65	.31	.70	.69
121314	162	.65	.29	.70	.66
311332	179	.64	.30	.70	.65
331143	177	.65	.28	.70	.60
124233	156	.65	.28	.70	.65
133431	175	.65	.28	.70	.66
113144	159	.64	.30	.70	.64
343211	173	.63	.30	.70	.63
314414	160	.62	.29	.65	.64
244341	162	.62	.28	.65	.63
231224	181	.62	.30	.65	.62
434112	167	.60	.29	.60	.65
343433	170	.60	.31	.65	.60
333333	165	.60	.30	.65	.60
442134	180	.59	.30	.60	.61
421441	183	.58	.31	.60	.59
413323	166	.58	.30	.55	.58
444444	169	.58	.31	.60	.56

Given the extremely low R^2 and the non-significant coefficients, the two first criteria of our model were not met. Therefore, we did not proceed to compare the TTO values with the predicted values.

Table 2 Regression coefficients for the 18 predictor dummy variables (levels across the six McSad domains)

McSad Domain attribute level (2-4)	**Unstandardized Coefficients**		
	B	**Std.Error**	**Sig.**
(Constant)	7.101	.128	.000
Emotion			
2	-.039	.117	.736
3	-.335	.122	.006
4	-.509	.120	.000
Self-appraisal			
2	-.136	.120	.258
3	-.156	.124	.209
4	-.194	.130	.136
Cognition			
2	.057	.118	.631
3	-.103	.116	.377
4	-.017	.123	.891
Physiology			
2	-.128	.118	.280
3	-.166	.120	.167
4	-.124	.119	.295
Behaviour			
2	-.313	.119	.008
3	-.160	.117	.173
4	-.430	.120	.000
Role-functioning			
2	.040	.117	.733
3	-.212	.118	.071
4	-.212	.124	.088

DISCUSSION

This paper reports on a study aiming to develop a scoring algorithm for the McSad depression-specific classification system. The McSad is recommended for valuations of depression, for which commonly used generic preference-based measures, such as the EQ-5D(Rabin & de Charro, 2001), as they have been found less sensitive (Papageorgiou, Vermeulen, Schroevers, Buskens, & Ranchor, 2013). That is because they include HRQoL domains not directly related to mental health. For this purpose we obtained the general population's TTO-based valuations for a total of thirty states generated by the McSad. We then modelled this data to develop a scoring algorithm that would predict values for all the 4096 states potentially generated by the McSad. Our predictive model consisted of variables describing levels across the six domains of the McSad classification system. Our purpose with such a value set was to enable indirect valuations of depression states and improve the validity of health state valuations in cost-utility analyses in depression interventions.

Overall, our attempts to develop the predictive scoring algorithm can only be considered as preliminary, as our predictive model was not satisfactory for developing a scoring algorithm. This was due to its poor performance in describing variability in our data and in the lack of statistically significant regression coefficients that would be used as weights in the scoring algorithm.

Nevertheless, we do consider that this study can provide valuable input for future studies aiming to develop scoring algorithms either for the McSad or for other depression-specific preference-based measures. For this reason, we focus below on reflecting on possible explanations for the poor performance of a model aiming to predict valuations of depression states, based on the description of these states (as reflected by levels in the domains).

The inability of the depression state description - as operationalized by the McSad levels in the six domains – to predict valuations for these states might raise some doubts concerning the descriptive properties of the McSad classification system. However, the properties of the McSad classification system have previously been demonstrated (Papageorgiou et al., 2013). Also, the McSad has been found valid in producing valuations (Bennett et al., 2000). Nevertheless, the fact that the McSad classification system generates a very large number of health states, as it includes more domains and levels than other common classification systems, and the fact that these domains may be mutually dependent hampers the development of a predictive model.

Alternatively, we can seek explanations for the poor predictive ability of our model, not in the description of the depression states, but in the quality of the values assigned to these states. We suggest that the limited range of valuations across the 30 states constituted the major challenge for the development of our predictive model. This limited range suggests that, on average, participants assigned similar values to depression states that described seemingly different levels of dysfunctioning across the different domains. In other words, there was a weak relation between the attached (perceived) values and the level of (described) severity of the depression state. Previously, a limited range of valuations for depression has been reported (Gerhards, Evers, Sabel, & Huibers, 2011), but a recent review reported a much larger range with an average (SD) value of 0.69 (0.14) for mild depression and of 0.25 (0.15) for severe depression (Mohiuddin & Payne, 2014), indicating that valuations for the more severe states can be considerably lower. Below we discuss possible explanations for the observed weak association between the valuations and the description of depression with regard to the methodology used for their elicitation and/or the mental health-specific nature of valuations.

Our methodology to elicit valuations was based on the TTO task. Although the TTO is a standard valuation task, a number of choices have to be made with regard to the development of each exact protocol, for example concerning the time frame employed or the upper scale of the valuation (Arnesen & Trommald, 2005). Regarding the methodology used in this study to elicit valuations, the first remark is that our upper scale was "absence of depression", in contrast to "perfect health" used in other studies (Bennett et al., 2000). This means that other co-morbidities were not excluded and thus individuals might be less willing to trade years if no perfect health would be gained, which might explain the relatively low valuations for the milder states of depression. A second remark is that we did not include the option of negative values in valuations. Negative valuations are sometimes possible to indicate states that are worse than dead, which might appear reasonable for some respondents in case of more severe states of depression. This might have resulted in the less negative average valuations for the more severe depression states. However, the percentage of zero values - which indicate states considered equivalent to dead ranged from 4.2% to 9.5%, so even if these states would have been valued negatively, such a small percentage cannot be expected to have a major influence on average valuations. A third remark is that we have chosen to name the condition described in the vignette. Therefore it is possible that individuals value "Depression", rather than the actual description presented to them. The effect of the label in valuation of health states has been previously reported (Rowen, Brazier, Tsuchiya, Young, & Ibbotson, 2012). It could also be that in the case of a mental health condition, more than in somatic condition, factors other than the described burden of the condition determine TTO decisions. Previously, factors such as perceived susceptibility or consideration of significant others have been found to be taken into account by individuals when performing TTO tasks to value depression (Papageorgiou et al., 2014). Additionally, we have decided to refrain from using an iteration procedure to elicit valuations, which might have affected our results. A final remark is that we opted for an online administration, which might have resulted in participants performing the valuations task less conscientiously.

The limited range of values might also be related to the mental health specific nature of valuations. Depression was described based on domains such as emotion or self-appraisal and we asked individuals to imagine living in the described condition and value their HRQoL. Domains known to be fundamental to HRQoL, such as mobility or pain, were supposed to remain unaffected. This implies that only a specific part of overall general health is affected, which can explain why valuations are restricted to the upper part of the 0 to 1 scale. Even more importantly, this part of overall health which is affected concerns mental health, whereas the unaffected part concerns somatic health. In valuations of somatic conditions participants are asked to trade somatically ill years to gain somatic health. In mental health valuations thought, although not explicitly stated, implicitly participants are asked to trade somatically healthy years to gain mental health. Interestingly, a previous study suggests that the general population might be less willing to pay to avoid mental illnesses, compared to somatic illnesses, even if the latter were considered less burdensome (Smith, Damschroder, Kim, & Ubel, 2012). Similarly, participants might have been more reluctant to trade years for mental health and this can contribute to our finding of less severe valuations.

Overall, the remaining question is whether the limited range of values is related to the methodology used in this study, and thus could be improved in future studies? Or are our result related to the mental health nature of valuations, i.e. that individuals underestimate the burden imposed by a mental health conditions, given that somatic health remains unaffected. The latter constitutes a challenge for future similar studies aiming to develop scoring algorithms for depression-specific preference-based measures? Currently, it is difficult to draw any definite conclusions.

It is not safe to exclude the option that the poor predictive power of our model is related to the selected model to describe this data. We have used a simple additive model, found to perform adequately in previous studies of similar type (Sundaram et al., 2010). In contrast to multiplicative models, simple additive models treat the contribution of each domain to the value as independent, thus ignoring potential interactions. Given that such interactions cannot be ruled out, it is possible that this partially contributed to the poor performance of our model. We could have used other types of more complicated models, for example by including participant characteristics, or combining levels, or by incorporating fixed and random effects models that would take into account that more than one point of our outcome correspond to the same participant. However, we chose not to proceed with any more complicated analyses. The reason is that the very poor performance of the simple additive model indicates that the input data were not appropriate for further analysis, due to the limited range discussed above. Overall, we acknowledge that this study did not achieve the aim of producing a scoring algorithm for a depression-specific preference-based measure. We propose that the raised challenges were mostly related to the limited range of values among the different states of depression. Above, we discussed possible reasons for this limited range, either conceptual or due to the methodology employed, and we believe this to be valuable knowledge for future studies with the same aim. It is therefore important to acknowledge the strengths of this study. The large sample of this study, representative of the general Dutch population, constitutes one of the main strengths of this study. Related to this is that we had multiple observations per state available. We also used a validated classification system (Papageorgiou et al., 2013) and a pilot-tested valuation elicitation method (Papageorgiou et al., 2014).

Future studies can build on this one and develop a scoring algorithm for the McSad classification system. One of the issues to be first explored could be how accurately the general population can assess the burden of depression. This would provide some indications on whether a sensitive valuation system could be yielded by a general population sample. This is particularly important as the general population is recommended for resource allocation decision making. Another issue that requires further investigation is whether the challenges related to the limited range of values would be present when a different methodology would have been used, for example, different states for direct valuations, a different administration TTO protocol or method, for instance, discrete choice modelling (Krabbe et al., 2014). With this study we made a first step in the developments of a scoring algorithm for a depression-specific preference-based measure. Future studies should follow up on this, as there is the potential to improve decision making with regard to depression related interventions, which may lead to more health benefits for people who experience depression.

REFERENCES

Arnesen, T., & Trommald, M. (2005). Are QALYs based on time trade-off comparable? - A systematic review of TTO methodologies. *Health Economics, 14*(1), 39-53.

Barrett, B., Byford, S., & Knapp, M. (2005). Evidence of cost-effective treatments for depression: A systematic review. *Journal of Affective Disorders, 84*(1), 1-13.

Bennett, K. J., Torrance, G. W., Boyle, M. H., & Guscott, R. (2000). Cost-utility analysis in depression: The McSad utility measure for depression health states. *Psychiatric Services, 51*(9), 1171-1176.

Bennett, K. J., Torrance, G. W., Boyle, M. H., Guscott, R., & Moran, L. A. (2000). Development and testing of a utility measure for major, unipolar depression (McSad). *Quality of Life Research, 9*(1), 109-120.

Brazier, J., Ratcliffe, J., Salomon, J., & Tsuchiya, A. (2007). *Measuring and valuing health benefits for economic evaluation*. Oxford: Oxford University Press.

Brazier, J., & Fitzpatrick, R. (2002). Measures of health-related quality of life in an imperfect world: A comment on dowie. *Health Economics, 11*(1), 17-19.

Brazier, J., Czoski-Murray, C., Roberts, J., Brown, M., Symonds, T., & Kelleher, C. (2008). Estimation of a preference-based index from a condition-specific measure: The king's health questionnaire. *Medical Decision Making, 28*(1), 113-126.

Dolan, P. (2002). The measurement of health related quality of life for use in resource allocation in health care. In N. J. Culer AJ (Ed.), *Handbook of health economics*. Elsevier Science.

Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. J., & Stoddart, G. L. (2005). *Methods for the economic evaluation of health care programmes*. Oxford: Oxford Medical Publications.

Gerhards, S. A. H., Evers, S. M. A. A., Sabel, P. W. M., & Huibers, M. J. H. (2011). Discrepancy in rating health-related quality of life of depression between patient and general population. *Quality of Life Research, 20*(2), 273-279.

Gold, M. R., Siegel, J. E., Russel, L. B., & et al. (1996). *Cost-effectiveness in health and medicine*. New York: Oxford University Press.

Green, C., Brazier, J., & Deverill, M. (2000a). Valuing health-related quality of life - A review of health state valuation techniques. *PharmacoEconomics, 17*(2), 151-165.

Green, C., Brazier, J., & Deverill, M. (2000b). Valuing health-related quality of life - A review of health state valuation techniques. *PharmacoEconomics, 17*(2), 151-165.

Grimison, P. S., Simes, R. J., Hudson, H. M., & Stockler, M. R. (2009). Deriving a patient-based utility index from a cancer-specific quality of life questionnaire. *Value in Health, 12*(5), 800-807.

Koenig, H., Guenther, O. H., Angermeyer, M. C., & Roick, C. (2009). Utility assessment in patients with mental disorders validity and discriminative ability of the time trade-off method. *PharmacoEconomics, 27*(5), 405-419.

Krabbe, P. F. M., Devlin, N. J., Stolk, E. A., Shah, K. K., Oppe, M., van Hout, B., . . . Xie, F. (2014). Multinational evidence of the applicability and robustness of discrete choice modeling for deriving EQ-5D-5L health-state values. *Medical Care, 52*(11), 935-943.

Lamers, L. M., Groot, C. A. U., & Buijt, I. (2007). The use of disease-specific outcome measures in cost-utility analysis - the development of dutch societal preference weights for the FACT-L scale. *PharmacoEconomics, 25*(7), 591-603.

Mohiuddin, S., & Payne, K. (2014). Utility values for adults with unipolar depression: Systematic review and meta-analysis. *Medical Decision Making, 34*(5), 666-685.

Papageorgiou, K., Vermeulen, K. M., Leijten, F. R., Buskens, E., Ranchor, A. V., & Schroevers, M. J. (2014). Valuation of depression co-occurring with a somatic condition: Feasibility of the time trade-off task. *Health Expectations : e-publication ahead of print*

Papageorgiou, K., Vermeulen, K. M., Schroevers, M. J., Buskens, E., & Ranchor, A. V. (2013a). Testing the McSad depression specific classification system in patients with somatic conditions: Validity and performance. *Health and Quality of Life Outcomes, 11*, 125-7525-11-125.

Petrillo, J., & Cairns, J. (2011). Development of the EXACT-U: A preference-based measure to report COPD exacerbation utilities. *Value in Health, 14*(4), 546-554.

Rabin, R., & de Charro, F. (2001). EQ-5D: A measure of health status from the EuroQol group. *Annals of Medicine, 33*(5), 337-343.

Revicki, D. A., Leidy, N. K., Brennan-Diemer, F., Sorensen, S., & Togias, A. (1998). Integrating patient preferences into health outcomes assessment - the multiattribute asthma symptom utility index. *Chest, 114*(4), 998-1007.

Rowen, D., Brazier, J., Tsuchiya, A., Young, T., & Ibbotson, R. (2012). It's all in the name, or is it? the impact of labeling on health state values. *Medical Decision Making, 32*(1), 31-40.

Smith, D. M., Damschroder, L. J., Kim, S. Y. H., & Ubel, P. A. (2012). What's it worth? public willingness to pay to avoid mental illnesses compared with general medical illnesses. *Psychiatric Services, 63*(4), 319-324.

Sundaram, M., Smith, M. J., Revicki, D. A., Elswick, B., & Miller, L. (2009). Rasch analysis informed the development of a classification system for a diabetes-specific preference-based measure of health. *Journal of Clinical Epidemiology, 62*(8), 845-856.

Sundaram, M., Smith, M. J., Revicki, D. A., Miller, L., Madhavan, S., & Hobbs, G. (2010). Estimation of a valuation function for a diabetes mellitus-specific preference-based measure of health the diabetes utility index (R). *PharmacoEconomics, 28*(3), 201-216.

Survey Sampling International.Survey sampling international http://Www.surveysampling.nl/.