Fachbereich Erziehungswissenschaft und Psychologie
der Freien Universität Berlin

Constructing Subtests Using Ant Colony Optimization

Dissertation
Zur Erlangung des akademischen Grades
Doktor der Philosophie (Dr. phil.)

Vorgelegt von
Diplom-Psychologe

Martin Schultze

Berlin, 2017

Primary Advisor: Prof. Dr. Michael Eid
(Erstgutachter)

Secondary Advisor: Prof. Dr. Tobias Koch
(Zweitgutachter)

Date of Defense: April 21st, 2017
(Datum der Verteidigung)

Acknowledgements

First, I would like to thank my advisors for their unbelievable support throughout the years - prior, during, and hopefully after this thesis project. Many thanks to Michael Eid, for giving me the freedom to write about this topic and always being ready to give me advice, when I needed it. Equally, a great deal of gratitude to Tobias Koch, who encouraged me to turn my pet-project into a full thesis while we were still colleagues and was willing to become my advisor after leaving Berlin.

Of course, many thanks to my (current and former) colleagues at the AB Methoden und Evaluation and beyond for always having an open ear for my questions and lamentations. Finding great people in the hallways for a thoughtful discussion or an uplifting chat went a long way in making the long days and short nights so enjoyable.

I want to thank my friends and family for being there for me and for bringing joy to those weekends and nights I did not spend writing, working, or simulating. Thank you, Lisa, for making this journey with me.

Contents

List of Figures

List of Tables

List of Symbols

I_m	Number of components in facet m
Ω	Set of constraints
α	Degree of non-linearity of the influence of the pheromone
β	Degree of non-linearity of the influence of heuristic information
η	Heuristic information
$\mathcal{B}_u$	Sets of multimethod assessments
$\mathcal{C}$	Set of components, e.g. items
$\mathcal{Q}_h$	Partitions of unique facets
$\mathcal{R}_v$	Sets of repeated measures
$\mathcal{S}^*$	Subset of permissible solutions
$\mathcal{S}$	Set of viable solutions
ϕ	Deposited pheromone
$\phi^{\max}$	Upper pheromone limit
$\phi^{\min}$	Lower pheromone limit
ρ	Evaporation coefficient
a_m	Capacity of facet m
c	Component, e.g. item
f	Objective function
k	Ant

m	Facet of a scale
$p_{s^{gb}}$	Target probability of constructing global-best solution
s	A constructed solution
s^{gb}	Global-best solution
s^{ib}	Iteration-best solution
s^{opt}	Optimal solution
t	Iteration
t^{gb}	Iteration since the last s^{gb}
x_{im}	Binary variable indicating the selection of item i to facet m

CHAPTER 1

Introduction

The assessment of constructs via questionnaires has a long tradition in psychological research. For example, Gault (1907) provides an overview of the history of questionnaire-based assessment in 1907, indicating a long standing tradition even then. And despite the numerous discussions on advantages and disadvantages of the questionnaire as an assessment method, it is still widely utilized in psychological assessment and research today. Especially closed-response questionnaires have been at the center of many psychological investigations, even when they are combined with other assessment methods, such as behavioral observation or, more recently, imaging techniques.

The reliance on questionnaires has spawned a great deal of guidelines, best-practices, and gold-standards for the construction of a psychological scale (e.g Clark & Watson, 1995; Dawis, 1987; Furr, 2011; Simms, 2008). A large number of these classical standards for scale construction are based on the trifecta of (*a*) the substantive component, (*b*) the structural component, and (*c*) the external component, as proposed by Loevinger (1957) in her seminal paper on scale construction. Roughly put, the three components can be thought of as corresponding to the three large phases in scale construction, though none of the three phases should be completely void of considerations regarding the other two components.

The first phase of scale construction is mainly concentrated on the substantive component and thus aimed at maximizing content validity. Content validity concerns the degree to which the contents of a scale represent the entirety of behavior which they are intended to allow conclusions about (Cronbach & Meehl, 1955). In this sense, its absence leads to wrong conclusions about the relation of the intended construct to other constructs and behaviors (e.g. Raykov & Marcoulides, 2011), and may thus jeopardize claims about any other type of validity. Loevinger (1957) extended this concept to substantive validity, to ensure the emphasis on inclusion of the broadest reasonable area of content into a scale. This is further differentiated by Allen and Yen

(1979) into face validity and logical validity, with the former indicating the extent to which the relation of a scale to its intended construct is obvious, and the latter indicating a more thorough definition of a constructs and the behaviors it is manifested in, as well as its integration into a more general nomological net. While these aspects are of critical importance to a scale's quality, Allen and Yen (1979, p. 96) note that "the determination of this type of validity is more subject to error than are other types of validity", further complicating this part of scale conceptualization. The focal point of the substantive phase is the generation of a pool of potentially useful items from theory and knowledge about the psychological concepts involved in the constructs which the scale is bound to assess. This means that it includes the strict definition of the construct and its manifestations in behavior, as well as an operationalization of the assessment of these manifestations. In this phase it is important to cover all possible areas which are part of the intended scope of the scale sufficiently, even at the expense of risking inclusion of irrelevant items and facets.

The second phase revolves around the structural component, aiming at selecting items from the pool in such a way that the scale fulfills a number of criteria about the desired structure. The focal points during applications are often the reliability and adherence to the theoretically derived factorial structure, be they classic unidimensionality or more complex structures. As formulated by Loevinger (1957, p. 664), items should be selected "on the basis of empirical properties, in particular, that those items be selected which best conform to an appropriate structural model". It is, however, of critical importance to consider that the empirical properties do not equate to statistical criteria alone, but may also encompass other experiences made in the application of the initial item pool. For example, items which were originally deemed acceptable by the experts constructing the item pool, may be unclear to respondents, resulting in confusion and non-response.

Finally, the third phase is focused on the validation of the scale on external criteria. These external criteria may either by assessed simultaneously, thus being indicators of concurrent validity, or be available only at a later time, thus making their predictability by the scale indicators of predictive validity. Concurrent validity is often subdivided further into convergent and discriminant validity, with the idea being that a scale should show significant overlap (convergence) with other indicators of the same construct or indicators of similar constructs and be appreciably different from indicators of different constructs.

Table 1.1 gives an overview of this extremely brief introduction into the general phases of scale construction. The focus in this thesis lies on the second phase, more specifically on the step of item selection. A great deal of attention has been given to aspects regarding item selection, because they are often deemed generalizable across a multitude of settings, allowing for the development of techniques independently of the construction of specific scales.

There are some basic contradicting interests inherent to any item selection, the most prominent of which is associated with the reliance on measures of internal consistency as indicators of

Table 1.1: Overview of the phases of scale construction.

Phase	Core Steps	Goal
Substantive	• Construct definition • Literature Review • Item formulation • Response format selection	General item pool
Structural	• Data collection • Assessment of item properties • Item selection	Potential scale
External	• Data collection • Assessment of predictive and concurrent validity • Cross-Validation of item selection	Scale validation

measurement reliability. Clark and Watson (1995) refer to this as the attenuation paradox, meaning that there is a paradox in the fact that increasing internal consistency becomes detrimental to a scale's construct validity after a certain point. Boyle (1991) goes even further, referring to internal consistency as a misnomer, indicating that it simply relates to redundancy of the assessed items. This, in turn, means that scales with high internal consistencies may simply assess a very narrow, specific facet of the much wider, originally intended construct. Thus, when operationalizing reliability as the consistency of multiple items in the same scale, there is a conflict in the balance between high reliability (in order to make "clean" assessments of a construct) and high substantive validity as put forth by Loevinger (1957). Falling too far into either extreme can render a scale unusable as it may either become too noisy and broad or too narrow and clean. This is, however, related only to the concept of reliability as internal consistency. The pure definition of reliability in classical test theory as the proportion of variance that is due to true-score differences, does not imply this conflict.

Eid and Schmidt (2014) describe six criteria, which should guide every item selection. These criteria are shown in Table 1.2 and relate (mostly) to qualities of the final selection, not the items themselves. Beyond these criteria guiding the empirical part of item selection, Eid and Schmidt (2014) explicate two additional steps in item selection: (a) utilizing expert ratings of an item's adequacy in terms of substantive validity and (b) using process analysis to determine the acceptance and understanding of items by subjects. As noted above, Loevinger (1957) believes the former part of the construction of the item pool, not part of the item selection, showing how closely these phases of scale construction are intertwined.

While stated relatively simply in Table 1.2, the criteria summarized by Eid and Schmidt (2014) have potentially far-reaching implications. For example, the criterion of test fairness implies that there are no consistent cultural effects on the assessment, thereby implying that any scale used for assessing people from different cultural backgrounds must be, to some extent,

Table 1.2: Criteria guiding item selection, according to Eid and Schmidt (2014).

Criterion	Description
Adherence to measurement model	Select only those items adhering to the proposed theoretical dimensionality of the scale.
Accuracy	Select items which increase the accuracy of the estimation of the latent construct. This is related, but not limited, to the reliability of an item.
Economy	Construct a scale only as resource consuming as is strictly necessary.
Reasonability	The measurement process is not unreasonably difficult or stressful for participants.
Fairness	The scale results do not systematically discriminate against certain groups.
Integrity	Results cannot be manipulated by the participant.

measurement invariant. This can, however, be an extremely difficult task to achieve. Some constructs may be more relevant in one culture than in another, thus making the representation much more fine-grained, necessitating more specific sub-domains in a questionnaire.

Thus, to allow for the potential fulfillment of the criteria, the assumption must be made that items with the relevant properties exist in the original pool of items. Note, that this assumption can be wrong for a wide array of psychological constructs and that there is potentially no way of fulfilling the criteria set towards good item selection when the item pool is inadequate. In such cases it would be necessary to either expand the item pool, thus also expanding the original definition of the content area of the scale, or the exact opposite, reduce the assessment to those sub-domains for which appropriate quality can be achieved.

In recent years especially, the criterion of economy has become more central in psychological research. Due to increasing focus on studies concerned with a multitude of interrelations between different constructs, longitudinal studies, ambulatory assessments, and online assessments, the economy of each single scale is often decisive to the inclusion of an entire construct in a study. This, in turn, has lead to wide usage of short scales. Kruyen, Emons, and Sijtsma (2013) reviewed six psychological journals between 2005 and 2010 and found 7.48% to report the use shortened scales. Despite this development having been met with a substantial degree of criticism (e.g. Kruyen, Emons, & Sijtsma, 2012; Kruyen, 2012; Kruyen et al., 2013; Smith, McCarthy, & Anderson, 2000), the usage of short-scales in psychological research has steadily increased (Ziegler, Kemper, & Kruyen, 2014). To minimize shortcomings of the reduced versions of scales a multitude of guidelines and clarifications have been proposed (e.g. Marsh, Ellis, Parada, Richards, & Heubeck, 2005; Stanton et al., 2002; Ziegler et al., 2014).

However different the perception, the basic problem in scale shortening is the same as it is in any form of psychometrically sound scale construction. Looking at the stages of scale construction as proposed by Loevinger (1957) and summarized in Table 1.1, the only notable difference is in the origin of the item pool during the substantive phase. While a full scale construction requires the explication of the psychological constructs and the nomological net surrounding it, this step has already been performed in a situation in which an existing scale is shortened. The original scale is equated to an original item pool and the remaining steps must be performed from there on. The very first of the nine specific sins of short form development presented by Smith et al. (2000) is the insufficiently validated original, indicating that even the first steps of scale construction cannot be disregarded in short-form construction. Instead of deriving the substantive validity of the item pool, an investigator shortening a scale must investigate the substantive validity of the original scale. The remaining steps must also be followed in short-form construction, just as they would be in the construction of a new, full scale. As Smith et al. (2000, p. 103) point out "many investigators assume that all of the reliability and validity evidence of the original, full-length measure applies automatically to the abbreviated version. This is false."

Because of this fundamental equality of short-form and original scale construction, they are not principally differentiated in this thesis. Instead, strategies to select items in the initial construction of a scale and those designed for short-scale construction are generally deemed suitable for both situations and the process is simply called "item selection".

The criteria shown in Table 1.2 all relate to internal properties of the scale. While Loevinger (1957) also explicitly separates the item selection (within the substantive phase) from scale validation (within the external phase), there is good reason to include external criteria into the process of item selection. In fact, some scales are defined primarily in relation to external criteria (e.g. scales used in selection processes of potential employees or scales used to assess recidivism risk in clinical psychology), making item selection based purely on internal criteria nonsensical. Instead, criteria often linked to the external phase - predictive, concurrent, convergent, and discriminant validity - may not only be relevant in evaluating a scale, but also be of central importance in item selection.

The following section will provide a brief overview over classical and modern approaches to item selection, before the conceptual representation of item selection used in this thesis is presented. Finally, this chapter will close with an introduction into Ant Colony Optimization.

1.1 Methods of Item Selection

According to Dawis (1987), classic techniques of item selection strategies can be separated into four distinct categories:

1. stimulus-centered methods,
2. subject-centered methods,
3. response scale methods, and

4. external criterion methods.

Stimulus-centered scales are based on an individual application, meaning that they aim at assessing the scaling of items within a single individual. These item can relate to a vast array of things - e.g. relationships, specific life events, fear inducing pictures - the relevance of the items is the meaning to the assessed individual. Thus, stimulus-centered scales do not aim at comparing individuals with each other, but rather aim at comparing stimuli within the specific representation of each individual. Because of this, they require much different approaches of item selection than scales aimed at assessing interindividual differences. While many early approaches of item selection focus on stimulus-centered scales (e.g. the Thurstone method or Q-sort; cf. Dawis, 1987), these approaches are not of interest here. Within this thesis, the main focus lies on the item selection in subject-centered scales.

Perhaps the first widely recognized item selection technique for subject-centered item selection was presented by Likert (1932). Herein, those items are chosen from the item pool which best discriminate between the groups of participants, which score highest and lowest on the total sum score across all items in the item pool. As Dawis (1987) points out, this approach was later translated into selecting items based on the item-total correlation. Neill and Jackson (1970) evaluated seven different approaches utilizing this item-total correlation to select items, concluding all seven to lead to solutions of equivalent psychometrical quality.

Because these methods are limited to unidimensional scales, factor analytical approaches were adapted for the use in item selection. The earliest of these approaches are based in exploratory factor analysis (EFA) and aim at selecting those items most clearly associated with each of the identified facets. As is the case with methods based on item-total correlations, these approaches heavily favor items which assess subdimensions in a similar fashion, irrespective of their substantive importance. To remedy this critique, Dawis (1987) proposed the identification of anchor items followed by the elimination of items which show too low correlations with these anchors. Factor analytical strategies can then be applied to this reduced item pool.

The third group of item selection techniques is labeled response scale methods and is traditionally associated with item response theory (IRT). Because an in-depth introduction into IRT is beyond the scope of this thesis, the reader is referred to Embretson and Reise (2000) and Hambleton and Swaminathan (1985) for more information. Within IRT, items can be selected from the original pool on the basis of their discrimination and difficulty parameters or by evaluating the item-information-function, to achieve a scale that provides a reliable assessment for the desired range of the underlying dimension. In contrast to the two types described, this shifts the focus of item selection from the scale to the items: items are selected due to their ability in assessing the underlying construct.

Beyond IRT, response scale methods have also been used in item selection for scales constructed under classical test theory (CTT), which is the theory of measurement this thesis will focus on. As a result of the wide spread of the use of confirmatory factor analysis (CFA) in psychological science during the early years of the 21st century, response scale methods and the necessary modeling have become widely accessible to psychological researchers (for detailed introductions to CFA see, e.g. Brown, 2015; Kline, 2011; Schumacker & Lomax, 2016). CFA approaches allow - much like IRT approaches - for the definition of a theoretical measurement model, which is assessed in its fit to a specific data setting. Thus, in contrast to EFA-based techniques to item selection, they provide an inferential test as to whether the selected items confirm to the proposed internal structure of the scale. In fact, under specific circumstances IRT and CFA are proven to be equivalent (Takane & de Leeuw, 1987), leading many researchers and modern textbooks to differentiate between IRT and CFA measurement models mainly due to the item format (e.g. Eid & Schmidt, 2014; Furr, 2011; ten Holt, van Duijn, & Boomsma, 2010; Raykov & Marcoulides, 2011). In CFA, items are most often selected due to two criteria: their adherence to the proposed measurement model and some indicator of their reliability. The former is often operationalized via modification indices, which help identify the specific restrictions imposed by a measurement model that lead to a discrepancy between the model and the empirical evidence (e.g. Brown, 2015). This is comparable to techniques utilizing EFA, where items are removed if they show cross-loadings above a certain threshold, thus undermining unidimensionality (Dawis, 1987; Neill & Jackson, 1970). Reliabilities are often included more indirectly, by selecting those items with the highest standardized factor loadings. While this may seem counter-intuitive, these are not necessarily indicators of an item conforming to the measurement structure. Instead, when assuming that an item measures only one construct (irrespective of the number of constructs underlying a scale), the standardized factor loading is simply the square-root of its reliability. Thus, these loadings indicate the reliability of an item given the assumed measurement model is true.

In a literature review of articles presenting the construction of psychological scales by ten Holt et al. (2010), only 6 of 46 studies used approaches other than factor analytical or IRT-based approaches. Of the 40, only ten used techniques not classically categorized as a response scale method, instead relying solely on EFA in item selection and validation of the measurement structure. Additionally, ten Holt et al. (2010) found that CFA approaches significantly outnumbered IRT approaches in scale construction, though only few of the studies explained their choice of one over the other.

The fourth large group of item selection methods are external criterion methods. As the name suggests, these techniques are mainly concerned with the relation of items to external criteria, thus leading to a selection process which maximizes predictive and concurrent validity. Most often these criteria are correlations or regressions with variables outside the item pool, but techniques aimed at maximizing group differences are also commonly found in clinical psychology. The last is similar to the Likert approach described above, but differs in that the groups for which differences in the scale are maximized are not derived from the item pool but are rather given by some external criterion (e.g. a clinical diagnosis). Correlative and regression approaches can be applied either at the level of the scale or at the item-level (Goldberg, 1972), with quite different results. Regression methods on the item-level use an external criterion as the dependent variable in multiple regression and add items from the item pool until the quality of the prediction no longer increases. As can be derived from the properties of multiple regression, this leads to a very broad, heterogeneous scale. Goldberg (1972) concluded that these methods resulted in scales that were not better - in terms of concurrent validity - than those arrived at solely by theory-guided item selection.

Notably absent from Dawis (1987) classification are methods selecting items based on their theoretical and substantive merit. This is due not to a proposed ignorance of theoretical criteria during item selection, but rather due to the fact that these criteria are less readily formalized and must therefore be defined on an application specific basis. Additionally, many ad-hoc impositions are often placed on the item selection process (e.g. including a similar amount of positively and negatively phrased items).

1.2 Strategies of Item Selection

The last section presented *techniques* to select items. Each single technique is easily implemented with modern analysis software, but often a scale is not constructed to fulfill only one criterion. Instead, multiple conflicting criteria define the quality of a scale. A scale must simultaneously be economically short and contain enough items to be reliable, be fair to all but provide enough intentional discrimination on a latent construct, be internally consistent and heterogeneous in order to predict outcomes well. These conflicts are also reflected in item selection, where it becomes necessary to utilize multiple techniques in selection.

Perhaps the most prominent conflict is between internal consistency and narrowness of the measured construct. Choosing items based solely on item-scale correlations or maximal factor loadings may (and most probably will) result in the assessment of what Boyle (1991) calls a “bloated specific”, meaning that one specific component of a more broadly defined constructed is exaggerated in its importance. While this may be avoided to some extent by appropriately defining sub-dimensions of constructs, these bloated specifics may also be the result of similarities in item formulation that are far removed from the original construct. For example, selecting items to represent the gregariousness sub-dimension of extraversion may be distorted by musical preferences, when asking about party activity. A remedy proposed by Dawis (1987) is the selection of anchor items and the elimination of items which do not correlated strongly enough with these anchors. This imposes a two-step strategy, which is symbolic of many selection strategies in scale construction because it implies a sequential form of item selection to account for different demands placed upon the final scale.

Stanton et al. (2002) propose the ten-step best-practice procedure for item selection presented in Table 1.3. These steps are created as a guideline for item selection and require the implemen-

Table 1.3: Ten-Step item selection procedure proposed by Stanton et al. (2002, p. 188).

Step	Goal	Method
1	Item level indices of external item quality	Item-criterion correlations
2	Indicators of internal item quality	Item-total correlations or factor loadings
3	Index of substantive item quality	Experts' or respondents' judgment of face validity
4	Item quality ranking	Three item quality judgments
5	Select high-rankings items	Quality ranking and professional judgment
6	Quality assessment of constructed scale	External correlations, correlation with full item pool
7	Reliability and internal consistency	Reliability and internal consistency estimates
8	Scale validation	Assess scale and external criteria in new sample
9	Quality assessment of scale	Repeat steps 6 and 7 in new sample
10	Cross-validate scale performance	Multiple Group SEM to compare scale in both samples

tation of a number of the techniques described in the previous section. The first four steps are intended to accommodate different demands posited towards a scale, thus combining different aspects of item quality in the selection of items.

A more rigorous approach was used by Johnson (2014) in the construction of the IPIP-NEO-120, a 120-item scale assessing the Big Five personality traits, each composed of six specific facets. He first eliminated items based solely on the item-total correlations and then evaluated the substantive quality of the resulting scale. Item combinations resulting in too narrow scales were exchanged for more meaningful items and reliability was evaluated again. In their item selection for a short-form of the Self-Description Qeustionnaire II (a questionnaire assessing the general self-concept of adolescents) Marsh et al. (2005) followed six self-imposed guidelines pertaining to item-total correlations, minimal cross-loadings and residual correlations, minimal non-response, and what they call “subjective evaluations of the content” (Marsh et al., 2005, p. 85). In their construction of the PPPM-SF - an observational checklist for the assessment of postoperative pain in children - von Baeyer, Chambers, and Eakins (2011) also utilized item-total correlations and item-criterion correlations.

Common to these selection strategies (and those implemented in countless other scale constructions in the past years) are three closely related shortcomings. First, they aim at fulfilling the criteria presented in the previous section for the constructed *scale*, but often select items based on information about the *items*. While there is no doubt that those qualities will be related, it is not necessarily the case that selecting items which show good qualities will result in a good scale. Selecting items without regard for the other items selected in the scale may result in very narrow scales, because very similar items will often have very similar qualities. Thus, the second shortcoming ensues: sequential item selection. In most cases items will be selected based on some sort of order - e.g. the ten-step procedure proposed by Stanton et al. (2002) proposes selecting items based on a ranking. This often results in the following problem: assume a scale assesses relationally aggressive behavior with items describing concrete behaviors and a rating scale, indicating how frequently a participant believes to display such behavior. Assume two items pertain to behaviors displayed towards family members. Because the scale is intended to assess behaviors from a wide array of contexts it may be reasonable to select the better performing of the two family-related items and then move on to selecting items describing behaviors towards friends, colleagues, etc. However it is quite possible that the item which showed the "worse" quality may result in a better scale, because it displays smaller residual correlations when combined with the items assessing behavior towards friends. To account for this, most approaches include a re-appraisal step where items are discarded if the potential for such problems is detected (e.g. Johnson, 2014). This leads to the third shortcoming: limited traceability of the entire item selection procedure. Running through the loop of selection, re-appraisal, elimination, selection, re-appraisal, and so on, results in selection criteria that are often not comprehensible to anyone not directly involved in the item selection process itself. In fact, this procedure is often so complicated that authors refrain from reporting it all (see Kruyen, 2012, for a critique).

These procedures become even more complicated when the requirements regarding scale qualities become more complicated. Selecting items with the goal of constructing a scale that is invariant across multiple populations often indicates performing these procedures either on the entire sample and assessing measurement invariance only in the re-appraisal step or performing the entire procedure separately in each sample and selecting those items which consistently show promising qualities. Both approaches have severe limitations, in that the first does not include group-specific information in the selection process itself, while the second has the potential to become incredibly convoluted when including many groups. The same is true when constructing scales designed for use in longitudinal settings, e.g. when assessing the development of constructs over the course of an intervention or when using scales in ambulatory assessment. Here too, information guiding item selection is available at multiple points and integrating the amount of information can become impossible.

Recent developments in item-selection strategy have attempted to alleviate this problem by using computer algorithms to select items from the pool of items (e.g. Danner et al., 2016; Janssen, Schultze, & Grötsch, 2015; Leite, Huang, & Marcoulides, 2008; Olaru, Witthöft, & Wilhelm, 2015; Schroeders, Wilhelm, & Olaru, 2015). In all of these approaches, measurement models and scale lengths are defined prior to item selection and items are then selected in line with predefined criteria. Leite et al. (2008) use the regression weight of the latent variable underlying their scale predicting a distal outcome as a selection criterion, while Janssen et al.

(2015) utilized the latent correlation structure of the original item pool as a criterion. Both of these, as well as Olaru et al. (2015) utilized model fit criteria to determine the adequacy of the measurement model with the given selection. Comparisons have shown selection strategies based on these algorithmic procedures to outperform classical approaches (e.g. Janssen et al., 2015; Olaru et al., 2015).

Underlying these current procedures is a different understanding of the problem posited by item selection, though none of the articles has clearly stated and described this conceptualization. The next section will present item selection as a combinatorial optimization problem and is intended to close this gap.

1.3 Item Selection as a Combinatorial Optimization Problem

As pointed out in the previous section, classical approaches to item selection often select items based on their specific qualities. However, the criteria this selection is aimed at achieving are mostly imposed on the final scale, not the items. In order to link the process of item selection more directly to its goal, it may be beneficial to focus on evaluating a set of selected items as a whole. Thus, the focus of item selection shifts from selecting good items to selecting items that are good when used together. This indicates a different approach to item selection, in the sense that the underlying problem that needs to be solved is conceptualized in quite a different fashion. When selecting items based on the merits of the items themselves, the problem is mainly in determining the quality of items and selecting the best. Quality, in this case, can refer to a broad selection of psychometrical properties of items, such as reliability, external validity, or content validity, among others. When instead focusing on the final selection of items, the problem becomes a combinatorial optimization problem.

This section will give a brief introduction to combinatorial optimization problems and their representation, before describing item selection as a specific problem-type of this class - the knapsack problem.

1.3.1 Combinatorial Optimization Problems

Though combinatorial optimization is a young field in terms of scientific investigation, early mathematical formulations of problems which can be subsumed under the class combinatorial optimization problems date back as far as 1784 (cf. Schrijver, 2005). In accordance to Blum and Roli (2003), combinatorial optimization problems can be defined as the search in a set for an object which best fulfills a given optimization criterion. While this definition may seem overly global it shows directly how many problems can be conceived of as being combinatorial optimization problems. In a more specific way of phrasing these problems, they can be seen as the search for an order or a subset of a discrete set of objects under given constraints, optimizing a quality function (Hoos & Stützle, 2004). Thus, the set which is being searched contains elements which can be combined or ordered in many different ways and the optimization problem is finding the *optimal* order or combination of elements.

In line with Dorigo and Stützle (2004) the more formal definition is given by the triple $(\mathcal{S}, f, \Omega)$. Here $\mathcal{S}$ represents the set of possible solutions, f is some objective function mapping the quality of a solution s to a non-negative real number $\mathbb{R}^{\geq 0}$, and Ω is a set of constraints.

Given an initial set of components $\mathcal{C} = \{c_1, c_2, ..., c_i, ..., c_I\}$, the set of possible solutions $\mathcal{S}$ is defined as the set of sequences $(c_a, c_b, ...)$ which fulfill the constraints provided by Ω. Additionally, there exists a non-empty set of optimal solutions $\mathcal{S}^{opt} \subseteq \mathcal{S}$, which contains at least one solution providing the optimum value for f.

Translated to the specific problem of item selection, $\mathcal{C}$ is the set of all items in the original item pool and $\mathcal{S}$ is the set of all possible combinations of items fulfilling the set of constraints provided by Ω. An example for a possible constraint ω is limiting the number of items which can be selected. The objective function can map the quality of the constructed scale by any number of quality indicators, such as reliability of the scale or the model fit of the measurement model. The framework of problem representation shown throughout this and the following section is intentionally general and imposes very little limitations on the possible definitions of $\mathcal{C}$, $\mathcal{S}$, f, and Ω, so as to allow a wide array of problems to be formulated within this framework.

Because the bulk of literature investigating the combinatorial optimization problems comes from the tradition of economics and resource management (cf. Hamann, 2015), f is often conceptualized as a cost function, making minimizing f the goal of optimization. In this thesis f will be seen as a quality function, making it necessary to maximize f. While this is of no consequence to the validity of the definitions, results, and conclusions of previous studies for this thesis, it should be noted explicitly, to avoid confusion regarding some definitions and equations.

With this very general and broad definition in place, a variety of problems can be seen as combinatorial optimization problems. One of the most studied subclasses of combinatorial optimization problems are so-called routing problems. As the name suggests, routing problems are concerned with finding optimal routes, passing through a number of pre-defined places. The most well-known among these routing problems is the traveling salesman problem (TSP), which was explicitly formulated during the middle of 19th century (cf. Schrijver, 2005), but finds its roots in many everyday applications. In a TSP the goal is to find the shortest (or most cost-effective) route through all cities a salesman must visit, ending at the original point of departure - a Hamiltonian cycle. Thus, for the TSP the set of possible solutions $\mathcal{S}$ is simply the full set of permutations on the set $\mathcal{C}$ of the initial components - i.e. $\mathcal{C}$ contains all cities and $\mathcal{S}$ contains all possible orderings of visiting them, under the constraints provided by Ω. The objective function is a cost function f, which is most often defined as the sum of traveling costs along the route. Finally, the set of constraints is $\Omega = \{\omega_1, \omega_2\}$, with ω_1 stating that a solution must begin and end at c_1 and ω_2 stating that all solutions in $\mathcal{S}$ contain each $c_i \in \mathcal{C}$ for all $i \neq 1$ exactly once.

Another well known subclass of combinatorial optimization problems are assignment problems, among which the quadratic assignment problem (QAP) is perhaps the most prevalent. This problem was first formulated by Koopmans and Beckmann (1955) and is discussed in great detail by Burkard, Çela, Pardalos, and Pitsoulis (1999). The main objective in a QAP is the assignment of a set of facilities to a set of locations while optimizing for the cost of necessary exchange between the facilities. This cost is defined as the product of distance and necessary flow between two facilities. While this problem was originally formulated for the placement of industrial complexes, it has been transposed to a number of different contexts, such as keyboard layout (Burkard & Offermann, 1977). Within the QAP the set of possible solutions $\mathcal{S}$ is therefore the set of all combinations of facilities and locations. The objective function is a cost function defined as the aforementioned product of distance and flow. The set of constraints is given by $\Omega = \{\omega_1\}$, where ω_1 states that the assignment of facilities to locations must be a bijection.

Both of these prominent representatives of combinatorial optimization problems have one specific thing in common: the number of solutions contained in $\mathcal{S}$ grows at a factorial rate - i.e. given n objects, the number of possible solutions is $n!$. Brute-force algorithms are not globally suitable for solving combinatorial optimization problems, because their runtime is bounded by the number of possible solutions: if the route of the TSP is to be *optimal* it must be shown that the route is better than all other possible routes. Thus, using brute force, all possible combinations would need to be evaluated.

The worst-case runtime (or, alternatively, the upper bound on runtime) of an algorithm is often denoted using the $\mathcal{O}(\cdot)$ notation (see Papadimitriou, 2003, for a concise introduction). This notation allows for an approximation of the time an algorithm can be expected to require, under the worst circumstances, in a shorthand fashion, because it allows ignoring lower-level terms. For example, if the runtime of an algorithm is denoted $\mathcal{O}(n^2)$ on a problem size n, the more accurate upper-bound may actually be given by $g(79n^2 + 12n + 100)$, for example. Thus, for small n the worst-case running time can be much larger than n^2, because constants and sub-quadratic terms are hidden in $\mathcal{O}(\cdot)$ notation. However, asymptotically the $\mathcal{O}(\cdot)$ notation will provide an estimate of worst-case running time. In practice, three functional classes of $\mathcal{O}(\cdot)$ are of extreme importance: (a) polynomial time, where the runtime depends on the problem size by some polynomial function, (b) pseudopolynomial time, where the runtime depends polynomially on more than one parameter of problem size, and (c) non-polynomial time, where the runtime exceeds polynomial functions of problem size - often meaning that some exponential function relates runtime to problem size. This type of notation is helpful, because it can be used to determine which types of algorithms are useful for solving problems of different complexity.

In the case of the TSP and the QAP, brute-force algorithms would be expected to run in $\mathcal{O}(n!)$ time. It should be noted, that more efficient exact algorithms have been proposed for most prominent combinatorial optimization problems - for the TSP: Baldacci, Hadjiconstantinou, and

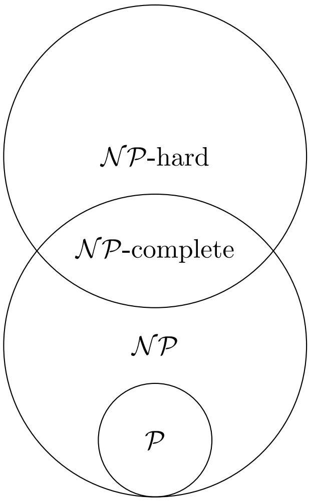


Figure 1.1: Euler diagram of complexity classes.

Mingozzi (2003), for an overview see Laporte (1992) and for the QAP: Ahmed (2013); Christofides and Benavent (1989); Mautor and Roucairol (1994), though none of them have been able to deliver solutions in polynomial time.

In addition to the fact that, for the described problems, around $\mathcal{O}(n!)$ time is required for the optimal solution s^{opt} to be found with exact and deterministic algorithms, it is also not easy to determine whether a solution is truly optimal, because optimality is defined only in relation to *all* other possible solutions in $\mathcal{S}$. This places these problems in the class of $\mathcal{NP}$-hard problems. Figure 1.1 gives an overview of algorithmic complexity classes as described by Reingold, Nievergelt, and Deo (1977). Classes become increasingly computationally complex towards the top of the Figure 1.1.

The "easiest" class is $\mathcal{P}$, which is a class of problems that are known to be solvable in polynomial time by deterministic algorithms - i.e. they have $\mathcal{O}(n^k)$ runtime with some constant value for k and problem size n. This means that the number of operations required to solve problems in $\mathcal{P}$ are bounded by a polynomial function of the size of the input. Next is the class of $\mathcal{NP}$ problems for which a solution can be verified in polynomial time. A problem in $\mathcal{NP}$ may require exponential time to be solved (using brute force), but whether a solution is correct or not can be verified "quickly". Therefore, though they may be hard to attain, solutions are not hard to check. A great deal of research has been conducted into the question of whether $\mathcal{P} = \mathcal{NP}$, which would mean that all problems for which a solution can be checked in polynomial time, the best solution can also be found in polynomial time. A definitive answer has not been found until now, in fact the $\mathcal{P}$ vs $\mathcal{NP}$ problem is of such importance that it is currently one of seven Millennium Problems for which a solution is prized at $1 million (Clay Mathematics Institute, 2000).

When a problem is harder than all other problems in $\mathcal{NP}$, but it still satisfies the condition of solutions being verifiable in polynomial time, it is considered $\mathcal{NP}$-complete. For $\mathcal{NP}$-hard problems, on the other hand, it is not necessarily the case that they can be verified in polynomial time.

For the purposes of this dissertation the main relevance is in the classification of problems as $\mathcal{NP}$-hard. Specifically, the importance is in the fact that if there exists no known deterministic algorithm to solve a specific $\mathcal{NP}$-hard problem in adequate time, then it becomes necessary to approximate a solution in adequate time using heuristic algorithms. A class of such heuristic algorithms is discussed in Section 1.4. To determine which complexity class the problem of item selection belongs to, the next section will present it as a specific variation on a well-known class of problems, for which many properties, such as the complexity class, are known.

1.3.2 Item Selection as a Knapsack Problem

Knapsack problems (KP) are a specific class of combinatorial optimization problems concerned with finding a subset s in a set $\mathcal{C}$, such that the subset s fulfills the optimality condition of some objective function f, while not exceeding given restrictions (e.g. Kellerer, Pferschy, & Pisinger, 2004). The term "knapsack" refers to the metaphor of packing a knapsack (or rucksack) before going on a hike. Each item you could pack has a certain value (e.g. water, food, and a raincoat all have their merits) but each item also weighs something. Because the knapsack (and your back) is limited in the weight it can carry, you can only pack things until a certain weight is reached. Each combination of items you could pack is a solution s and the set of possible solutions $\mathcal{S}$ is a set containing all combinations of components in $\mathcal{C}$. In line with Kellerer et al. (2004), the objective function in prototypical KP is given by

$$f(s) = \sum_{i=1}^{I} b_i x_i \tag{1.1}$$

where i is the running index of the components c, x_i is a binary variable indicating whether $c_i \in s$, and b_i is some benefit associated with selecting c_i. In classical cases describing KP, the benefit is often monetary but in the case of item selection it could be that an item benefits the scale by increasing reliability, for example. Additionally, the main constraint in Ω associated with KP is:

$$a \geq \sum_{i=1}^{I} w_i x_i \tag{1.2}$$

with each component c_i weighing w_i, and a representing the maximum weight which can be carried. In terms of item selection the simple case is given by setting a to the number of items that should be selected and setting all $w_i = 1$. But this does not necessarily have to be the case. When utilizing vignettes in assessment (e.g. Finch, 1987), it might be more sensible to define the weight of an item by the time it takes to complete it, because the length of vignettes can vary drastically and the time required for the entire questionnaire should not exceed certain boundaries.

It must be assumed that $a > w_i$ (all single weights are smaller than the capacity) and $a < \sum_{i=1}^{I} w_i$ (the sum of all weights exceeds capacity) to ensure non-triviality of the KP (Fréville, 2004). In the basic KP it can also be assumed that $b_i > 0$ and $w_i > 0$ (benefits and weights are strictly positive) because all instances not fulfilling these two criteria can easily be rewritten as instances fulfilling them (Kellerer et al., 2004). Equations (1.1) and (1.2) imply a major difference between the KP and many other combinatorial optimization problems: the order of c_i is irrelevant to the solution of the problem. Or, stated more formally:

$$(c_i, c_{i'}, ..., c_I) = (c_{i'}, c_i, ..., c_I) = \{c_i, c_{i'}, ..., c_I\}. \tag{1.3}$$

While the assumption of the irrelevance of order is not strictly necessary in item selection, it is sensible for two reasons: (a) in most cases the original item pool is not presented in multiple orders, indicating that conclusions drawn about the order of items from the process of item selection would not be grounded in data, and (b) in most CFA applications the order of indicators is inconsequential to indexes of model fit, estimates of reliability, and correlations between latent variables, because these models are at least globally covariance equivalent, meaning that they produce the same model-implied covariance matrices (c.f. Hershberger, 2013). In this case two solutions s and s' would generate $f(s) = f(s')$, making them functionally indistinguishable in terms of quality. Due to these two aspects, it is sensible to assume the order of items as irrelevant, because this greatly reduces the size of the search space - making the problem smaller and thus easier to solve.

With these constraints in place, a component can be selected independently of the actual components selected previously, once the sum of their weights and the capacity of the knapsack are controlled for:

$$p(x_i' = 1 | x_{i''}, \sum_{i \in s}^{I} w_i x_i, a) = p(x_i' = 1 | \sum_{i \in s}^{I} w_i x_i, a) \qquad \forall i' \neq i''. \tag{1.4}$$

In addition to Equations (1.3) and (1.4), the formulation of the basic KP implies that

$$(b_i | c_i \in s) = (b_i | c_i \in s') \tag{1.5}$$

meaning that the benefit of any component is independent of the solution under consideration, and therefore of the other components that are chosen alongside it. This is often the case in classical KP - e.g. when selecting which packages to load onto a delivery truck, because each package has an associated worth, irrespective of the worth of the other packages. Problems that fulfill these criteria (and are thus KP) can be solved in pseudopolynomial time and are proven to be $\mathcal{NP}$-complete (Kellerer et al., 2004).

In CFA-based item selection the assumption stated in Equation (1.5) restricts the selection criteria of items to trivial item characteristics such as means and variances. Because these characteristics are required to be independent of other items in s, finding an optimal selection of items based solely on these criteria does not require a CFA approach. The criteria used in CFA-based item selection are most often based on either fit of the measurement model or scale reliability and are, as such, dependent on the constellation of items investigated in any given model.

To avoid the restriction made in Equation (1.5), Gallo, Hammer, and Simeone (1980) introduced the quadratic knapsack problem (QKP). This differs from the basic KP in the definition of the objective function, which is

$$f(s) = \sum_{i=1}^{I} \sum_{i'=1}^{I} b_{ii'} x_i x_{i'}, \tag{1.6}$$

indicating that benefits of c_i are dependent upon the selection of $c_{i'}$. This implies a quadratic matrix of benefits $b_{ii'}$, instead of the vector of b_i assumed in the basic KP, hence the name *quadratic* knapsack problem. The overall benefit of selecting two components c_i and $c_{i'}$ is given by $b_{ii} + b_{i'i'} + b_{ii'} + b_{i'i}$. The four separate benefits denote the global benefit of selecting c_i, the global benefit of selecting $c_{i'}$, the increase in the benefit of c_i when coupled with $c_{i'}$, and the increase in the benefit of $c_{i'}$ when coupled with c_i. Thus, it is not necessarily the case that the benefit matrix is symmetric. QKP are known to be strongly $\mathcal{NP}$-hard, meaning that they are known to not be solvable even in pseudopolynomial time. This means that solutions to these problems must be found using heuristic algorithms or approximation algorithms (Pisinger, 2007).

The case of item selection utilizing CFAs goes beyond the QKP, because the benefits of a component c_i are not additively dependent on the vector of the other components selected for s. Instead the additional benefit of selecting c_i when having already selected $c_{i'}$ may depend on all other components selected and is thus possibly unique in every $s \in \mathcal{S}$, making it necessary to denote benefits as b_{is} - i.e. both component and solution specific. In CFA-based item selection this can be seen as the benefit an item has over an arbitrary replacement in terms of solution quality. This makes this approach to item selection an I-dimensional KP. Stated more formally, the I-dimensional KP can be viewed as:

$$f(s) = \sum_{i=1}^{I} b_{is} x_i \tag{1.7}$$

$$\text{subject to} \qquad a \geq \sum_{i=1}^{I} w_i x_i. \tag{1.8}$$

Note that, though the benefit of c_i is solution specific in Equation (1.7), the weight w_i in Equation (1.8) is not. As pointed out above, $w_i = 1$ can be used to make a the number of items selected in the final scale - this is independent of the constellation of items selected. Similarly, using other values of w_i to encode the time it takes to finish vignettes, the w_i should be approximately independent. Because the QKP is a special case of the I-dimensional KP and the QKP is strongly $\mathcal{NP}$-hard, the problem of CFA-based item selection formulated as an I-dimensional KP must also be strongly $\mathcal{NP}$-hard.

These elaborations on the KP and QKP in relation to item selection are only feasible in a special case: selecting items from a common pool of items to represent one single facet, i.e. when a scale is used to unidimensionally assess only one specific construct. Utilizing this approach to depict the problem of item selection in general requires two extensions: (a) the use of multiple knapsacks m to depict multiple facets and (b) selection restrictions indicating which item is intended to measure which facet(s).

The first extension is, quite sensibly, called the multiple knapsack problem (MKP). In these cases Equations (1.1) and (1.2) are extended to

$$f(s) = \sum_{m=1}^{M} \sum_{i=1}^{I} b_{im} x_{im} \tag{1.9}$$

$$\text{subject to} \qquad a_m \geq \sum_{i=1}^{I} w_{im} x_{im}. \tag{1.10}$$

Here, m indicates one of $(1, ..., M)$ knapsacks and benefits, as well as weights, become knapsack dependent. In terms of item selection this knapsack specificity is sensible, because items may well be more beneficial when chosen to measure one facet of a questionnaire, than when chosen to select another. Combining Equations (1.7) and (1.9) provides the objective function

$$f(s) = \sum_{m=1}^{M} \sum_{i=1}^{I} b_{ims} x_{im}. \tag{1.11}$$

This objective function is subject to the constraint provided by (1.10), because weights are not assumed to be solution specific, as shown in Equation (1.8).

The second necessary extension - the inclusion of assignment restrictions - was first presented

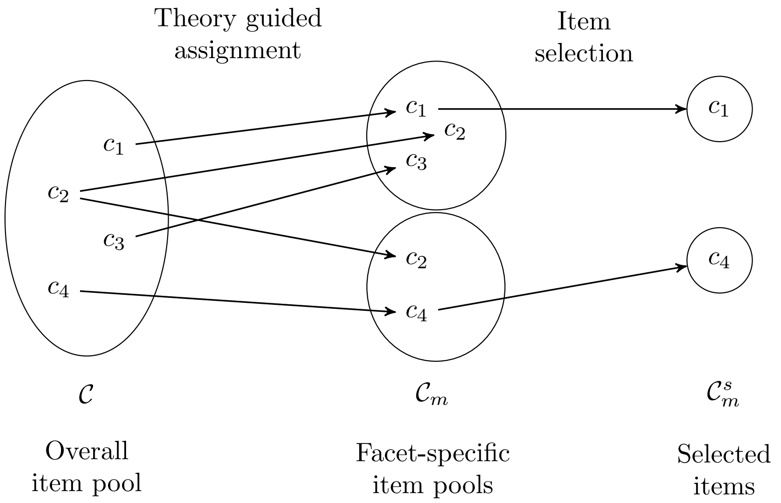


Figure 1.2: Overview of the element mapping in item selection when construed as a knapsack problem.

by Dawande, Kalagnanam, Keskinocak, Ravi, and Salman (2000) for regular MKP. There are two possibilities for imposing assignment restrictions. The first is via the definition of $\mathcal{M}_i \subseteq \mathcal{M}$, i.e. by defining specific subsets of knapsacks ($\mathcal{M}_i$) which are viable to hold component c_i. In this case restrictions are placed on each knapsack m, so that it can only contain specific components in $\mathcal{C}$. The other possibility is the definition of knapsack specific sets of components, viable to be assigned to m as $\mathcal{C}_m \subseteq \mathcal{C}$. Throughout this thesis the second variant will be used.

The definition of $\mathcal{C}_m \subseteq \mathcal{C}$ makes it necessary to impose two restrictions in addition to that given by Equation (1.10) (c.f. Dawande et al., 2000):

$$\mathcal{C}_m^s \subseteq \mathcal{C}_m \tag{1.12}$$

$$\mathcal{C}_m^s \cap \mathcal{C}_{m'}^s \equiv \emptyset \qquad \forall m \neq m' \tag{1.13}$$

Equation (1.12) states that the components assigned to knapsack m in solution s must be elements of the specific set of components eligible for assignment to knapsack m. Equation (1.13) states that for any solution s the sets of components assigned to two different knapsacks must be disjoint. MKP with assignment restrictions remain strongly $\mathcal{NP}$-hard (Dawande et al., 2000), therefore their extension to the case for I-dimensional MKP is also strongly $\mathcal{NP}$-hard. The restrictions imposed by Equations (1.12) and (1.13) can easily be integrated into the I-dimensional multiple knapsack problem, because neither relate directly to the benefits encoded in b_i.

To give a more direct indication of the relevance of these general definitions in the context of this thesis, the sets and restrictions can be related to the process of item selection. Figure 1.2 shows the implied mapping of elements in this process. $\mathcal{C}$ is the starting point, because it is the set of all items available. The items contained within this exhaustive set are denoted c_i, with $i \in (1, ..., I)$ and I indicating the total number of items. Derived from theory (or perhaps the dimensionality of the long version of a scale which is being shortened) there are M distinct facets which are measured by the scale, each of which is denoted by an $m \in (1, ..., M)$. A facet is measured by I_m items, which are contained in the set $\mathcal{C}_m$, which is, of course, a subset of the set containing all items ($\mathcal{C}$). Items may measure more than one facet, and can thus be elements of more than one $\mathcal{C}_m$, though this is generally not recommended under the criteria of CTT. Selecting items then means drawing items from the facet-specific subsets $\mathcal{C}_m$ and assigning them to the facet-specific subset $\mathcal{C}_m^s$ used in solution s. During this selection process three constraints $\Omega = \{\omega_1, \omega_2, \omega_3\}$ are imposed. ω_1 is given by Equation (1.10) and states that an item c_i may only be selected if its selection does not result in the sum of weights w_{im} of the items selected for facet m exceeding the capacity a_m of the facet. Specifically, if the number of items allowed to measure facet m is set to a_m and all $w_{im} = 1$, an item may only be selected as long as there are still "empty spots" in the short version. The second constraint, ω_2, is given by Equation (1.12) and states that items can only be selected to measure a facet m if it was stated, a priori, that it measures facet m. The final constraint, ω_3 as given by Equation (1.13), states that an item may only be selected once in any given solution. Thus, while items may be eligible to measure different facets in the original item pool, they must be assigned to measure a specific facet in the selection process.

1.4 Ant-Colony-Optimization

Ant-Colony Optimization (ACO) is a meta-heuristic for solving $\mathcal{NP}$-hard combinatorial problems on the basis problem solving behavior of certain types of ants. It was first described as a meta-heuristic by Marco Dorigo and colleagues (Dorigo & Di Caro, 1999a, 1999b; Dorigo, Di Caro, & Gambardella, 1999), but first algorithmic approaches subsumed under this meta-heuristic are quite a bit older (e.g. Colorni, Dorigo, & Maniezzo, 1991).

Meta-heuristics are somewhat ambiguously defined - so much so that Blum and Roli (2003, p. 270-271) cite four different definitions before listing nine global components common among them. The points of central importance to this thesis are that they (a) guide search processes via a set of strategies, (b) are not problem specific, and (c) are approximate and not deterministic. The last point is of crucial importance, because it emphasizes that meta-heuristics are designed to find "very high quality solutions to hard, practically relevant combinatorial optimization problems in a reasonable time" (Dorigo & Stützle, 2004, p. 25). Thus, meta-heuristic (and heuristic, for that matter) approaches do not provide a guarantee of finding optimal solutions to a problem. Instead, the focus lies on knowing that a problem is computationally demanding - most often $\mathcal{NP}$-hard

- and finding an appropriate strategy to approximate the optimal solution to a satisfying degree in an appropriate amount of time. This is done via stochastic optimization, i.e. using random variables, their distributions and their manipulation to guide the search in a promising direction (cf. Fouskakis & Draper, 2002).

Among these meta-heuristics, simulated annealing, evolutionary algorithms, and swarm algorithms are perhaps the most prominent. ACO is a representative of the last of these categories because it utilizes swarm behavior in its manipulation of search direction probabilities.

In this chapter the basic metaphor underlying ACO will be described - the emergent exploration behavior of ant colonies. Subsequently, the first and most influential ACO algorithm, the Ant-System, is described in Section 1.4.2. The following section then describes the $\mathcal{MAX} - \mathcal{MIN}$ Ant-System which is of central importance to this thesis because it is the algorithmic approach used to select items here. Section 1.4.4 will then give a brief overview of other approaches that have been developed within the context of the ACO metaheuristic as well as some other refinements often employed alongside one of these approaches.

1.4.1 Real Ants

As the name suggests, the real-life behavior of ants is the foundation of the ACO meta-heuristic. Ants live in societies of varying complexity and size, and are thus subsumed under the category “social insects” (a category that also includes, termites, bees, and wasps). While by no means consensus, a strong case has been made for the view of a colony as one super-organism instead of a gathering of multiple organisms (Hölldobler & Wilson, 1990). The degree of complexity of this organism has been focus of a substantial amount of research and a general taxonomy is provided by Anderson and McShea (2001). As Anderson and McShea (2001) note, the complexity of ant colonies can be described by a multitude of correlated dimensions with the general tendency toward more complex societies with more specialized individuals when the society is larger (see Kesebir, 2012, or Kennedy & Eberhart, 2001, for a discussion of the parallels to human society).

The aspect of ant societies that is of interest in ACO are the so-called foraging strategies. The greater extent of specialization of individuals in a complex society leads to a proportional increase in the amount of food a forager must secure, in order for the super-organism as a whole to survive. This implies that more complex ant societies must forage in a more efficient way than less complex societies do. As summarized by Beckers, Goss, Deneubourg, and Pasteels (1989) larger - and therefore more complex - ant societies rely on increasingly more intricate foraging behavior to secure the needed amount of food. While small and simple ant societies rely mainly on individual, non-cooperative foraging, larger societies rely on different types of recruitment strategies to achieve cooperation between foragers to ensure a higher degree of efficiency in the search for and gathering of food resources. As Anderson and McShea (2001, p. 223) point out “there is a shift from information processing by individuals to emergent properties of a set

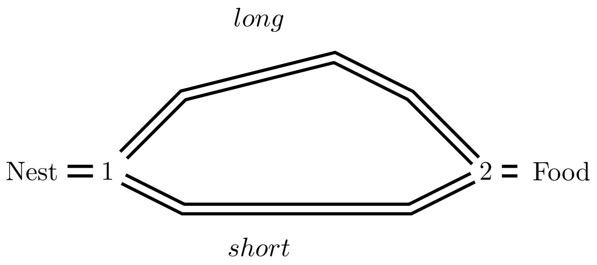


Figure 1.3: A schematic of the double-bridge experiment conducted by Goss et al. (1989).

of essentially probabilistically behaving individuals mediated through signals" with increasing societal complexity. This type of communication between foragers is called stigmergy - i.e. communication that is achieved by manipulations of the environment. In the case of ants this manipulation is most often achieved by the emission and detection of pheromones. It is important to note that this communication is not deterministic but probabilistic in its consequences for behavior.

A research group surrounding Jean-Louis Deneubourg (e.g. Deneubourg, Pasteels, & Verhaeghe, 1983; Goss et al., 1989; Deneubourg, Aron, Goss, & Pasteels, 1990) studied the foraging behavior of several species of ants under experimental conditions. All of the examined species of ants use a mass recruitment strategy to ensure cooperation between foragers, but especially the Argentine ant (*Iridomyrmex humilis*) is of relevance to ACO.

Figure 1.3 shows a schematic representation of the double-bridge experiment conducted by Goss et al. (1989) - aptly named, because ants can choose between two bridges, a short one and a long one (i.e. the set of possible choices is $\mathcal{L} = \{short, long\}$), both leading from their nest to a food source. The experimental setup was designed to ensure that the very first ant would favor neither choice, thus making $p(l) \approx .5$ for both paths $l \in \mathcal{L}$ for this ant. Having chosen either *short* or *long* at the first decision node, the ant reaches its goal and then returns to the nest carrying food. On this return it reaches decision node 2 and again determines probabilistically which route to choose. This choice is now, however, biased slightly towards choosing the path it previously took. The special thing about the Argentine ant is that - unlike most other ants - it deposits pheromones both while searching for and while returning from food resources. This pheromone encourages ants to follow its path - meaning that in the double bridge experiment the probability of choosing path l is influenced by the amount of ants that previously chose that same path.

Via Monte-Carlo-Simulations Deneubourg et al. (1990, p. 163) were able to recreate the behavior of ant colonies for a double-bridge experiment in which both branches are equal (i.e. $short = long$). They were able to show that the probability for selecting path l can be expressed as:

$$p(l) = \frac{(c + n_l)^\alpha}{(c + n_l)^\alpha + (c + n_{l'})^\alpha}. \tag{1.14}$$

Here, n_l represents the number of ants that previously chose l, c represents some constant attraction parameter of a choice that no ant has previously made, and α is the degree of non-linearity of the choice function.[1] This equation only holds in cases in which the set of possible choices contains only two elements. The value of c determines the degree with which the relative amount of ants that previously chose l affects the probability of the current ant to make the same choice and is assumed to be equal for all possible choices in $\mathcal{L}$. Using empirical evidence Deneubourg et al. (1990) determined $c \approx 20$ and $\alpha \approx 2$ to provide a realistic depiction of the foraging behavior of the Argentine ant in this experimental setting.

Equation (1.14) shows that, even in settings in which both alternatives are equal, ants will eventually use one of the two alternatives predominantly. Because the pheromones deposited by previous ants make a certain decision more probable, more ants will make the same decision and deposit pheromone making the same choice even more probable and so on. This autocatalytic effect means that $\lim_{n\to\infty} p(l) = 1$ and $\lim_{n\to\infty} p(l') = 0$ for the case with two mutually exclusive choices l and l'. This means that eventually one of the paths (l) will be chosen exclusively, irrespective of the fact that both are of equal length, because n - the number of ants that made the journey from nest to food resource - approaches infinity given enough time.

This behavior has no real purpose in cases with two equally viable choices. However, Goss et al. (1989) portrayed this behavior in cases in which one bridge is r-times longer than the other (the case depicted in Figure 1.3). When one path is shorter than the other ($r \neq 1$) pheromones will accumulate faster on the shorter path, because more ants can travel along the short path in the same amount of time. This increases the probability of choosing the *short* path over time, which means that food will be collected more efficiently overall.

Goss et al. (1989) generalized the equation shown in (1.14) so that the weighting of a chosen path is not only dependent on the pure number of ants that previously chose it, but instead on the amount of pheromone deposited on a path l:

$$p(l) = \frac{(c + \phi_l)^\alpha}{(c + \phi_l)^\alpha + (c + \phi_{l'})^\alpha}. \tag{1.15}$$

The amount of deposited pheromone is a function of the number of ants that made a specific choice, the quality of the chosen path (in this case its length), as well as a degree of pheromone evaporation that simply occurs over time. Because the evaporation of pheromones is very slow in reality, Goss et al. (1989) determined that it has an ignorable influence on the amount of pheromone that is deposited, and thus on the system describing the foraging process as a whole, in short-term experimental conditions. For the later use in ACO, however, pheromone evaporation

[1] Please note, that the notation used here does not correspond to the original notation used by Deneubourg et al. (1990), but rather to that used by Dorigo and Stützle (2004) to avoid confusion in later sections of this thesis.

plays a key role.

This depiction of the mass-recruitment strategy used by the Argentine ant focuses on the positive effects it has on the efficiency of food collection. When a food resource becomes available the colony will explore many possible ways to reach the resource, because each ant behaves probabilistically at each decision node. In the early stages of the search little pheromone has been deposited, making different choices at decision nodes more or less equally likely. Thus, there is a high degree of exploration during this phase of foraging. As time goes on, pheromones accumulate faster on those paths that are more efficient and more and more ants tend to use these paths, marking a high degree of exploitation in this phase of foraging. Thus, the super-organism as a whole explores a wide array of possibilities early, before focusing on those choices that lead to the best results - ideal properties for solving combinatorial problems. However, Goss et al. (1989) were also able to show, that once a path is established the occurrence of a better solution will not be noticed by the colony.

The basics of ants' foraging behavior described in this section are the foundation of the ACO metaheuristic. The core concepts that are important for the transfer from real ants to the artificial ants used in ACO are (a) the probabilistic behavior of each single ant in which (b) choice probabilities are influenced by pheromones which are (c) left by other ants making the same choice leading to (d) an autocatalytically increasing preference for better paths. The reason why ants' foraging behavior is so interesting for general-purpose problem solving is the concept of emergence. Emergence describes the idea that simple entities following simple rules can create highly complex systems, which is one of the core ideas of swarm intelligence in general (c.f. Kennedy & Eberhart, 2001). In this context, this phenomenon allows ant colonies to solve complicated problems (i.e. foraging) by only providing simple rules to a large enough number of ants and relying on the emergence of complex problem solving strategies from the number of ants as a whole.

1.4.2 Ant System

The first adaptation of the behavior of ants, as described in the previous section, to algorithmic problem-solving was the Ant System (AS) approach proposed by Marco Dorigo and colleagues (Colorni et al., 1991; Dorigo, 1992). While it is possible to construe algorithms that are simpler and closer to the real-world behavior of ants (e.g. the appropriately named Simple-ACO shown by Dorigo & Di Caro, 1999b; Dorigo & Stützle, 2004, pp. 11-21) AS is historically the first algorithm that is subsumed under the ACO metaheuristic.

As with most ACO algorithms, AS was designed with respect to a construction graph G representing the problem. Much like in the double-bridge experiment described in Section 1.4.1 there are paths l_{ij} that connect nodes i and j within this construction graph. However, different from the experiment conducted by Goss et al. (1989) these connections l_{ij} are not different paths between the same nodes, but rather connections between different nodes ($i \neq j$).

In AS the probability of choosing to go from i to j at iteration t is given by

$$p(i,j|t) = \frac{[\phi_{ij}(t)]^{\alpha}[\eta_{ij}]^{\beta}}{\sum\limits_{j=1}^{J}[\phi_{ij}(t)]^{\alpha}[\eta_{ij}]^{\beta}}. \tag{1.16}$$

This is very much in line with the equation for choice probabilities determined by Goss et al. (1989), but incorporates two important extensions beyond Equation (1.15): it is not limited to choosing different paths to get to the same destination and it incorporates heuristic information (η) into determining choice-probabilities. In the original article this heuristic information consisted of a simple inverse of the euclidean distance that needed to be traveled if path l_{ij} were to be chosen, but it allows for the inclusion of any prior information on the quality of connections. Equation (1.16) additionally implies that this prior information is the same throughout the entire algorithmic process, while the pheromone $\phi_{ij}(t)$ is a function of the iteration step t (sometimes confusingly referred to as time). From t to $t+1$ these pheromones are updated by

$$\phi_{ij}(t+1) = \rho\phi_{ij}(t) + \Delta\phi_{ij}(t,t+1) \tag{1.17}$$

with ρ representing the evaporation coefficient. As pointed out in Section 1.4.1, Goss et al. (1989) determined that under experimental conditions the evaporation of pheromones was irrelevant to the problem solving behavior of real ants. In AS, however, ρ is crucial to obtaining solutions, because it guarantees that all choices that are not made, are forgotten over time, making them even less likely in the future.

$\Delta\phi_{ij}(t,t+1)$ is simply the pheromone added to a chosen path in iteration t. Colorni et al. (1991) proposed three different approaches (ANT-density, ANT-quantity, and ANT-cycle) that differ in how and when $\Delta\phi_{ij}(t,t+1)$ is computed. The first two differ in their computation: ANT-quantity allocates a constant pheromone to all paths that are part of the constructed solution, while in ANT-density this constant is divided by the cost (the distance traveled along path l_{ij}). In both cases the pheromone update is made after each choice, while in ANT-cycle it is made only after the entire solution has been constructed. In the case of ANT-cycle the constant pheromone is divided by the cost of the entire solution:

$$\Delta\phi_{ij}(t,t+1) = \begin{cases} \frac{c}{ll_s}, & \text{if } l_{ij} \in s_k \\ 0, & \text{else} \end{cases} \tag{1.18}$$

where c is an arbitrary positive constant and ll_s is the total length of the solution constructed by ant k. In applications to 10- to 75-city TSPs, Colorni et al. (1991) found the ANT-cycle to be superior to the other two approaches, which is why it is the only one of the three discussed in more detail here. In the following AS will always refer to the ANT-cycle AS approach, unless

Algorithm 1: AS (ANT-Cycle) as proposed by Colorni et al. (1991)

```
Require: $\mathbf{H}$, $\alpha$, $\beta$, $\rho$, $T$, $K$
procedure AS($\mathcal{S}, f, \Omega$)
    set arbitrary $\phi_{ij}(0) = \phi_{i'j'}(0)$
    $f(s^{gb}) \leftarrow 0$
    randomize all $k$ ants starting position
    while $t \leq T$ do
        if $\mathcal{L}_k$ is not empty then
            choose $l_{ij}$ using (1.16) for each $k$
            remove $l_{ij}$ from $\mathcal{L}_k$
        else
            compute $f(s_k)$
            determine $\Delta\phi_{ij}(t, t+1)$ with (1.18)
            compute $\phi_{ij}(t+1)$ with (1.17)
            if $f(s_k) > f(s^{gb})$ then
                $s^{gb} \leftarrow s_k$
            end if
            reset all $\mathcal{L}_k$
            randomize all $k$ ants starting position
            $t \leftarrow t + 1$
        end if
    end while
    return $s^{gb}$
end procedure
Result: $s^{gb}$
```

explicitly stated otherwise.

Algorithm 1 shows the AS algorithm in simplified pseudo-code. The arguments that are required for AS to find a viable solution are the components of Equation (1.16) - the non-linearity coefficients α and β as well as the heuristic information $\mathbf{H}$ - to determine the probability of any given choice, the evaporation coefficient ρ used in the pheromone update shown in Equation (1.17), the number of iterations T, and the number of ants per iteration K.

Given these parameters, the AS procedure is then applied to the optimization problem $(\mathcal{S}, f, \Omega)$. As discussed in Section 1.3 this problem is characterized by the set of viable solutions $\mathcal{S}$, the pheromone function f, mapping the quality of a solution s to $\mathbb{R}^{\geq 0}$, and a set of constraints Ω.[2]

AS is initialized at some arbitrary pheromone level that is the same for all choices at all decision nodes. Thus, the initial solution is biased in its choices only by the heuristic information provided via $\mathbf{H}$. All ants k are placed on random starting locations in G.

All ants simultaneously choose a connection l_{ij} on the construction graph with the probabilities given by Equation (1.16) and move along their individual paths. The path an ant k has

[2] Equation (1.18) shows the pheromone function used by Colorni et al. (1991) in their original article. As noted in Section 1.3.1 these functions are often construed as *cost* functions, meaning that they need to be minimized. In the case discussed here f is considered a *quality* function, meaning that it needs to maximized.

just taken is removed from its set of possible paths $\mathcal{L}_k$. This is done until the set of possible paths is empty, at which point each ant has constructed a complete solution s_k. After each of these iterations, the pheromones of the paths are updated via Equation (1.18), thus changing the probabilities of the choices in the next iteration. The quality of the solution s_k is assessed by some objective function f and the globally best solution s^{gb} is memorized. The globally best (gb) solution is characterized as the solution with max $f(s)$. In the original articles by Colorni et al. (1991) and Dorigo (1992) the objective function is given by $\frac{c}{ll_s}$ - as shown in Equation (1.18) - with c being some arbitrary constant and ll_s being the length of the final path. However, in theory, this objective function can be any function describing the quality of the solution. This process continues until the predefined number of iterations T is reached.

As mentioned in Section 1.4.1, this process autocatalytically improves the probabilities of making choices that were made before. The pheromone function originally used by Colorni et al. (1991) ensures that choices which lead to a shorter total distance of s are updated with more pheromone, if they are made, than choices that lead to a longer total distance. Given enough iterations this should theoretically result in a solution that is close to optimal. The result of the algorithm is the globally best solution s^{gb} that was best found during the T iterations.

The influence of the parameters α, β, c, and ρ has been thoroughly investigated (e.g. Colorni et al., 1991; Dorigo, Maniezzo, & Colorni, 1996; Dorigo & Stützle, 2004). All simulations have shown that AS can be used to obtain very good solutions in relatively short time, if these parameters are chosen correctly. In general, it has been found that (a) increases in α and β lead to more exploitation, but less exploration, (b) heuristic information is necessary for AS to settle into promising areas of the search space, (c) ρ reaches a saddle point for which it ensures best results, and (d) c is irrelevant to the performance of AS. Additionally, Colorni et al. (1991) were able to show that if a $K \leq 16$ is chosen, AS is unable to find optimal solutions, irrespective of the chosen T.

Many adaptations of AS have been proposed since its original presentation by Colorni et al. (1991). Three specific algorithms - Elitist AS (Dorigo, 1992; Dorigo et al., 1996), rank-based AS (Bullnheimer, Hartl, & Strauß, 1997), and $\mathcal{MAX}$-$\mathcal{MIN}$ AS (Stützle, 1998) - extended the original AS by incorporating the concept of elitism. As the name suggests, elitism is simply the practice of favoring the solutions provided by some ants over the solutions provided by others. The Elitist AS extends the normal AS by an additional pheromone update for those solutions that are the best (i.e. those that have the shortest path ll_s). In rank-based AS, the solutions found by the ants of one iteration step t are sorted by their path length ll_s and a weighting parameter is applied to Equation (1.18) which is proportional to the inverse of this rank. Both of these procedures lead to a faster convergence on good solutions - ensuring more exploitation at the expense of exploration. The $\mathcal{MAX}$-$\mathcal{MIN}$ AS is discussed in more detail in the next section. Other general extensions of AS include the Ant Colony System (Dorigo & Gambardella, 1997) and the Approximate Nondeterministic Tree-Search (Maniezzo, 1999), both of which will be discussed in a bit more detail in Section 1.4.4. Beyond these general algorithms, numerous problem-specific AS extensions have been proposed (see Dorigo & Stützle, 2004, for an overview).

1.4.3 $\mathcal{MAX}$-$\mathcal{MIN}$ Ant System

The $\mathcal{MAX}$-$\mathcal{MIN}$ Ant System ($\mathcal{MM}$AS, Stützle, 1998; Stützle & Hoos, 2000) is an extensively tested and well performing pure-ACO extension of AS (extensions that go beyond the ACO framework are discussed in Section 1.4.4). Beyond the addition of elitism as briefly touched upon in Section 1.4.2, it also introduces limits to the pheromone trails to ensure that the probability of choosing any viable connection l_{ij} is never 0 or 1, thus guaranteeing some minimal exploration throughout the entire search procedure.

Stützle and Hoos (2000) propose two different types of elitism. In the first variant, iteration-best solutions s^{ib} are used for the pheromone update, meaning that, for every iteration t, only the one solution with $\max f(s)$ deposits pheromone on the paths, while the results of the remaining $K - 1$ ants are discarded. As shown in Algorithm 1, one of the basic principles of the ANT-Cycle version of AS is that K ants are initialized during each iteration t and that pheromones are updated only after each iteration. This means that selection probabilities for these K ants within one single iteration are equal. $\mathcal{MM}$AS - in contrast to AS - uses only the selection made in s^{ib} during this iteration t to deposit pheromones. The second variant goes even further by discarding all solutions that are not the globally best solution s^{gb} found at iteration t, meaning that a solution must be better than all preceding solutions to deposit pheromones. This means that the computation of $\Delta\phi_{ij}(t, t+1)$ changes to

$$\Delta\phi_{ij}(t, t+1) = \begin{cases} f(s^{ib}), & \text{if } l_{ij} \in s^{ib} \\ 0, & \text{else} \end{cases} \tag{1.19}$$

or

$$\Delta\phi_{ij}(t, t+1) = \begin{cases} f(s^{gb}), & \text{if } l_{ij} \in s^{gb} \\ 0, & \text{else} \end{cases} \tag{1.20}$$

respectively. In both equations $f(s)$ indicates the quality of the solution s (e.g. c/ll_s in Equation (1.18) for AS). This means that all ants k that did not construct either an iteration-best or a global-best solution have no influence on the pheromones - and are thus inconsequential for the choices made in the future.

In addition to elitism, $\mathcal{MM}$AS introduces upper and lower bounds that are imposed upon $\phi_{ij}(t)$. Equations (1.17) and (1.18) show that in AS, with any $\rho < 1$ the pheromone of a choice l_{ij} is limited to

$$\lim_{t\to\infty} \phi_{ij}(t) = \frac{f(s)}{1-\rho} \tag{1.21}$$

Thus, the pheromone of choices which are only made in solutions with zero quality ($f[s] = 0$) will approach zero as t becomes increasingly large. On the other hand the pheromones of choices that are part of an optimal solution are bounded proportional to the maximum quality (i.e. $f[s^{opt}]$) that can be achieved.

When AS has run for a large number of iterations, the differences in pheromones will become increasingly large. This is partly wanted, because it ensures that better choices are more probable in future iterations, but it may also lead to stagnation - i.e. the state in which the same solution is found exclusively because the probabilities of some choices are (practically) 0 while those of others are (practically) 1. Such a state may occur even though the global-best solution s^{gb} is far below optimal. In $\mathcal{MM}$AS this stagnation is avoided by ensuring that the difference between the probabilities of two alternative choices does not become too large.

To ensure this, two new parameters $\phi^{\min}$ and $\phi^{\max}$ are introduced. The upper limit of the pheromones is given by

$$\phi^{\max} = \frac{f(s^{gb})}{1-\rho}, \tag{1.22}$$

which is an estimate for the upper limit of the pheromones, as given in Equation (1.21), that is obtained by using the global-best solution s^{gb} as a placeholder for the optimal solution s^{opt}. Because a new s^{gb} can be found at any t, the upper pheromone limit $\phi^{\max}$ becomes dependent upon the iteration - making it a dynamic upper limit.

The definition of an adequate lower limit $\phi^{\min}$ is not quite as simple as that of the upper limit, but simulations have shown the lower limit to be of greater importance to the performance of $\mathcal{MM}$AS (Stützle, 1998; Stützle & Hoos, 2000). Stützle and Hoos (2000) propose computing $\phi^{\min}$ by

$$\phi^{\min} = \frac{\phi^{\max}(1 - \sqrt[n]{p_{s^{gb}}})}{(avg - 1)\sqrt[n]{p_{s^{gb}}}}, \tag{1.23}$$

where $p_{s^{gb}}$ is the desired probability of an ant constructing the global-best solution s^{gb} after the search has ended, n is the total number of choices that need to made, and avg is the average number of possibilities at each choice. The $p_{s^{gb}}$ must be defined for each application separately, because it should depend on the size of $\mathcal{S}$ - in cases in which the set of possible solutions is small, $p_{s^{gb}}$ should be quite large. The influence of $p_{s^{gb}}$ on the performance of $\mathcal{MM}$AS will be discussed in more detail in Section 3.1.7.

Using both of these limits, choice probabilities will no longer tend towards 0 and 1, but rather towards

$$\lim_{t\to\infty} p_t(i,j) = \frac{\phi^{\min}}{\sum\limits_{j=1}^{J} \phi_{ij}(t)} \tag{1.24}$$

for bad choices, and towards

$$\lim_{t\to\infty} p_t(i,j) = \frac{\phi^{\max}}{\sum\limits_{j=1}^{J} \phi_{ij}(t)} \tag{1.25}$$

for good choices. This implies that as long as $0 < \phi^{min} \leq \phi^{max} < \infty$ a path l_{ij} will never be selected - or not selected - with certainty based solely on the information provided by the pheromones. The addition of certain heuristic information via η_{ij} may have this effect, however.

To ensure that the pheromones stay within the bounds that were just shown, Stützle (1998) define an extension of Equation (1.17) to

$$\phi_{ij}(t+1) = \begin{cases} \phi^{\min}, & \text{if } \rho\phi_{ij}(t) + \Delta\phi_{ij}(t,t+1) < \phi^{\min} \\ \phi^{\max}, & \text{if } \rho\phi_{ij}(t) + \Delta\phi_{ij}(t,t+1) > \phi^{\max} \\ \rho\phi_{ij}(t) + \Delta\phi_{ij}(t,t+1), & \text{else} \end{cases} \tag{1.26}$$

which computes the new pheromone as ϕ^{max} if the pheromone computed by Equation (1.17) exceeds the the upper bound, ϕ^{min} if the result of (1.17) is below the lower bound, and the result of (1.17) in all other cases.

Besides preventing stagnation, the implementation of minimum and maximum pheromone levels allows for a straightforward definition of a convergence criterion. Stützle and Hoos (2000) define this criterion as

$$\forall i, j \neq j' \quad \exists! j : \qquad (\phi_{ij} = \phi^{\max}) \wedge \left(\phi_{ij'} = \phi^{\min}\right). \tag{1.27}$$

This means that $\mathcal{MM}$AS converges if, for every origin i there is one, and only one, choice j that has the upper pheromone limit $\phi^{\max}$ associated with it, while all other choices j' have the lower pheromone limit $\phi^{\min}$ associated with them. In this case the algorithm can be considered converged because there is one solution s^{gb} that is obviously favored above all other solutions. This is advantageous in comparison to AS, because (traditional) AS requires a fixed number of iterations to be run, regardless of the status of the actual search. This can lead to sub-par solutions, because T was chosen to be too small, or to a waste of time and resources, because T was chosen to be too large. In $\mathcal{MM}$AS, T can be used as an abort criterion if the algorithm cannot construct a viable solution or fails to converge for some other reason.

With these two extensions in place, $\mathcal{MM}$AS can be written in pseudo-code as shown in Algorithm 2. As was shown in Algorithm 1 for AS, $\mathcal{MM}$AS requires α, β, and $\mathbf{H}$ as arguments

Algorithm 2: $\mathcal{MM}$AS as proposed by Stützle (1998) with the iteration-best deposit rule.

```
Require: $\mathbf{H}$, $\alpha$, $\beta$, $\rho$, $T$, $K$, $p_{s^{gb}}$
procedure $\mathcal{MM}\text{AS}(\mathcal{S}, f, \Omega)$
    set $1 \ll \phi_{ij}(0) = \phi_{i'j'}(0)$
    $f(s^{gb}) \leftarrow 0$
    randomize all $k$ ants starting position
    while $t \leq T$ and (1.27) is not fulfilled do
        if $\mathcal{L}_k$ is not empty then
            choose $l_{ij}$ using (1.16) for each $k$
            update $\mathcal{L}_k$
        else
            compute $f(s_k)$
            $s^{ib} \leftarrow \arg\max f(s_k)$
            determine $\Delta\phi_{ij}(t, t+1)$ with (1.19)
            compute $\phi_{ij}(t+1)$ with (1.26)
            if $f(s^{ib}) > f(s^{gb})$ then
                $s^{gb} \leftarrow s^{ib}$
                compute $\phi^{\max}$ with (1.22)
                compute $\phi^{\min}$ with (1.23)
            end if
            reset all $\mathcal{L}_k$
            randomize all $k$ ants starting position
            $t \leftarrow t + 1$
        end if
    end while
    return $s^{gb}$
end procedure
Result: $s^{gb}$
```

to compute any given selection probability. Additionally, ρ is required for the pheromone update performed via Equation (1.26). The maximum number of iterations T is used as an abort criterion and the number of ants to be used per iteration is denoted K.

With these parameters in place the $\mathcal{MM}$AS procedure can be applied to the optimization problem $(\mathcal{S}, f, \Omega)$.

One important aspect that has not been considered so far is the initial pheromone level and its limits. Stützle and Hoos (2000) propose to initialize the pheromones to an arbitrarily large value - i.e. one that is larger than any reasonable limit that can be computed by Equation (1.21) with an idealized s^{opt}. This step is performed in line 2 of Algorithm 2 and results in a hard reset of all pheromones to $\phi^{\max}$ after the first iteration in line 13. The opposite - initializing all pheromones to a number that is much smaller than any reasonable $\phi^{\min}$ - has also been proposed, but was determined to lead to too little exploration and too fast convergence (Stützle, 1998; Stützle & Hoos, 2000).

After the initial value of the global best solution is initialized to 0 (line 3) and all initial starting positions are randomly assigned (line 4), the algorithm is set to run either until the predefined maximum number of iterations K has been evaluated or until the convergence criterion presented in Equation (1.27) is fulfilled.

During each iteration, each ant k iteratively chooses an l_{ij} and updates the set of possible choices $\mathcal{L}_k$ (because the ant is now in node j all viable paths now start in j). When the set of possible choices is empty the ant has constructed a solution s_k for which the pheromone value is computed (line 10). The iteration-best solution is determined and assigned to s^{ib}. Using $f(s^{ib})$ and Equation (1.19) the pheromone deposit for choices included in s^{ib} is determined and then inserted into Equation (1.26) to determine the pheromones of choices for iteration $t+1$ (lines 12 and 13).

To ensure the viability of ϕ^{max} and ϕ^{min} as limits of the pheromones, these are updated if a new global-best solution s^{gb} is found. As discussed earlier, these limits are theoretically dependent on the optimal solution s^{opt}, but because this solution is unknown, s^{gb} is used as an approximation of the optimal solution. Therefore, both ϕ^{max} and ϕ^{min} must be updated each time a new s^{gb} is found.

At the end of each iteration t, the set of viable choices $\mathcal{L}_k$ is reset, all ants are re-initialized to random starting positions, and the iteration counter is increased (lines 19 through 21). Once either $t = T$ or the convergence criterion is fulfilled the algorithm ends, returning the global-best solution s^{gb}.

Algorithm 3 shows the pseudo-code for $\mathcal{MM}$AS when allowing only ants constructing global-best solutions to deposit pheromone. In contrast to the approach depositing pheromone for each iteration-best solution, the s^{ib} is not stored and $\Delta\phi_{ij}(t, t+1)$ is the same until some $f(s_k) > f(s^{gb})$, as shown in Equation (1.20). Therefore, the pheromones on the choices l_{ij} are updated using the same procedures until the current best is surpassed, meaning that the choices

Algorithm 3: $\mathcal{MM}$AS as proposed by Stützle (1998) with the global-best deposit rule.

```
Require: $\mathbf{H}$, $\alpha$, $\beta$, $\rho$, $T$, $K$, $p_{s^{gb}}$
procedure $\mathcal{MM}\text{AS}(\mathcal{S}, f, \Omega)$
    set $1 \ll \phi_{ij}(0) = \phi_{i'j'}(0)$
    $f(s^{gb}) \leftarrow 0$
    randomize all $k$ ants starting position
    while $t \leq T$ and (1.27) is not fulfilled do
        if $\mathcal{L}_k$ is not empty then
            choose $l_{ij}$ using (1.16) for each $k$
            update $\mathcal{L}_k$
        else
            compute $f(s_k)$
            if $f(s_k) > f(s^{gb})$ then
                $s^{gb} \leftarrow s_k$
                compute $\phi^{\max}$ with (1.22)
                compute $\phi^{\min}$ with (1.23)
            end if
            determine $\Delta\phi_{ij}(t, t+1)$ with (1.20)
            compute $\phi_{ij}(t+1)$ with (1.26)
            reset all $\mathcal{L}_k$
            randomize all $k$ ants starting position
            $t \leftarrow t + 1$
        end if
    end while
    return $s^{gb}$
end procedure
Result: $s^{gb}$
```

in s^{gb} are reinforced after every iteration.

1.4.4 Other ACO Algorithms and Extensions

Beyond the two algorithms discussed in the two previous sections, a number of different ACO approaches have been proposed. In line with Dorigo and Stützle (2010, p. 239), Table 1.4 provides an overview of the most influential algorithms. This is by no means an exhaustive list - there are many problem specific adaptations and variants of ACO approaches - but it contains the core of ACO development for static problems. ACO extensions to dynamic problems, such as the AntNet (Di Caro & Dorigo, 1998), AntHocNet (Ducatelle, 2007), or the population based ACO (Guntsch & Middendorf, 2002), are not discussed here, because item selection constitutes a static problem.

Of the approaches presented in Table 1.4, Elitist AS, Rank-Based AS, $\mathcal{MM}$AS, and BWAS all represent straightforward extensions of the classical AS approach. In all cases, new concepts or parameters are introduced, but the core procedures of determining selection probability via Equation (1.16) and updating these choice probabilities directly by influencing the pheromones ϕ_{ij} remain intact. While $\mathcal{MM}$AS was discussed more thoroughly in Section 1.4.3, the Elitist AS and Rank-Based AS can be described quickly. In Elitist AS the classical AS is simply modified to not deposit each ants pheromone, but instead deposit only the pheromones associated with the s^{gb}, thereby reinforcing the strongest solution up until this point (Dorigo et al., 1991). The underlying idea is that by reinforcing this best solution, future ants are more likely to construct similar solutions, which are of higher quality. In Rank-Based AS, this elitist approach is combined with the notion of all ants depositing pheromone by allowing $K + 1$ ants per iteration to deposit pheromone, where the s^{gb} constitutes the "+1". These ants are ordered by their solution quality and their pheromone deposit is weighted by their rank - with better ranks receiving larger weights (Bullnheimer et al., 1997). The Best-Worst Ant System (BWAS; Cordón et al., 2002) extends the ideas of AS by adding to the inverse concept of elitism: evaporating the pheromones on choices made in the iteration-worst solution (if they do not also belong to the best solution).

The Ant Colony System (ACS; Dorigo & Gambardella, 1997) deviates from the tradition of extensions of the AS approach by introducing a pseudo-random rule to choice probabilities instead of the direct relation proposed in Equation (1.16). Without going into too much detail, this is achieved by imposing a doubly probabilistic choice. First - with a set probability - the choice with the maximum product of $\phi_{ij}\eta_{ij}^{\beta}$ is selected. If this choice is not made, Equation (1.16) is applied with $\alpha = 1$. One of the common properties of ACS and $\mathcal{MM}$AS is that both belong to a class of ACO algorithms for which convergence in value (i.e. finding the optimal value at least once) can be proven given infinite runtime (Dorigo & Stützle, 2004). While this does not seem like much of a feat, it is extremely valuable to know that a probabilistic algorithm will find the optimal solution given enough time and will not reiterate sub-optimal solutions infinitely.

Table 1.4: Overview of the most influential ACO Algorithms.

Name	Primary Citation	Properties
AS	Dorigo (1992)	• All ants deposit pheromone
Elitist AS	Dorigo, Maniezzo, and Colorni (1991)	• s^{gb} deposits after each iteration
Rank-Based AS	Bullnheimer et al. (1997)	• s^{gb} deposits pheromone • Ants deposit pheromone proportional to their rank in iteration
$\mathcal{MM}$AS	Stützle (1998)	• s^{gb} or s^{ib} deposit pheromone • Imposes limits on pheromones • Initializes pheromones to maximum
ACS	Dorigo and Gambardella (1997)	• Pseudorandom proportional rule for selection • Only choices in s^{gb} are updated • Ants remove pheromone with each choice
BWAS	Cordón, de Viana, and Herrera (2002)	• s^{gb} deposits pheromone • Only worst solutions evaporate • Pheromones are mutated with set probability • Includes reinitialization when stagnant
ANTS	Maniezzo (1999)	• Projects total $f(s)$ at each choice • Bases choice on projection of overall $f(s)$ • Uses one balance parameter instead of α and β
Beam-ACO	Blum (2005)	• Generates partial solutions and projects $f(s)$ • Extends partial solution to multiple partial solutions stepwise
ACOPD	P. Zhu, Zhao, and He (2010)	• Pheromones diffuse to close decisions

For the classical AS, for example, this is not the case and it is possible for any given run of AS to not construct the optimal solution, even when given infinite runtime. However, because both $\mathcal{MM}$AS and ACS impose lower limits on the pheromones, the probability of constructing any solution s never decreases to 0, but instead to a function of ϕ^{min}. Thus, while it possible for both, $\mathcal{MM}$AS and ACS, to find the optimal solution slower than even purely random search, it is impossible for either of them to not find it at all - again, if allowed to run infinitely long. This important feature makes them much more attractive than the other purely AS-based ACO approaches for the use in item selection, because there is a guarantee that increasing resources (time, in this case) has a benefit: it makes finding the optimal solution, at least once, more likely.

The ACS and $\mathcal{MM}$AS approaches are, perhaps, the most widely received and applied approaches within the ACO framework (Dinh, Mamun, & Dinh, 2005; Dorigo & Stützle, 2004; López-Ibáñez, Stützle, & Dorigo, 2015). Both perform similarly well under most conditions (Dorigo & Stützle, 2004; Escario, Jimenez, & Giron-Sierra, 2015; Fidanova, 2007; López-Ibáñez et al., 2015), with $\mathcal{MM}$AS being slightly more susceptible to the influences of parameter scheduling (Stützle et al., 2010), a topic discussed in Section 3.1.8. Both provide a very general, flexible framework, which can be applied to a wide array of problems.

The ACOPD is a modern extension of the AS approach that includes an additional mimicry of real-life ant foraging called pheromone diffusion (P. Zhu et al., 2010). In these approaches, neighborhoods which are close to choices with very high levels of deposited pheromone will also benefit from these pheromones. The intensity of this benefit is modeled in accordance to some model about the distance between nodes (Ji, Song, Liu, & Zhang, 2013).

Other modern ACO approaches often combine the AS strategy with another algorithmic approach to constructing solutions. The Elitist-Mutated AS combines the $\mathcal{MM}$AS approach with genetic algorithms (Afshar, 2009). The Approximate Nondeterministic Tree Search (ANTS; Maniezzo, 1999) combines ACO with branch and bound strategies (J. D. Little, Murty, Sweeney, & Karel, 1963), as does Beam-ACO (Blum, 2005). The basic idea behind branch and bound strategies is the construction of partial solutions, which are faster to construct and evaluate than entire solutions (branching), determining a bound on the quality of the partial solution (bounding), and eliminating branches which cannot lead to the optimal solution. It should be noted that these algorithms are exact, meaning that they lead to the optimal solution with certainty, but may require very long to complete. This is integrated into ANTS by eliminating choices that are known to not be feasible (because of their bounds) from the set of possible of choices at that point and choosing among the remainder via the ACO selection strategies. Beam-ACO differs from ANTS in that it implements Beam Search, a class of branch and bound inspired approximate algorithms, which construct multiple partial solutions, evaluate them, and continue with a fixed number of best partial solutions.

The problem with both of these approaches is that they require knowledge of the search space in some fashion and make it necessary to evaluate partial solutions at every step. As pointed out in Section 1.3.2, understanding item selection as an IMKAR does not allow for meaningful evaluation of partial solutions without additional assumptions, because solution quality is dependent upon all components in a solution.

One of the most successful extensions proposed for most ACO algorithms is local search (Bonabeau, Dorigo, & Theraulaz, 1999; Dorigo & Stützle, 2004). The idea in local search is that solutions constructed by the ACO algorithm are - often randomly - changed in very few locations repeatedly to determine the best solution within this specific neighborhood. This is done as an optimization step after a solution is constructed but before it is allowed to deposit pheromones. The main advantage of this process is that it is much cheaper, computationally speaking, to change constructed solutions in few locations than it is to construct new solutions. However, this approach is only truly advantageous if the construction of solutions is expensive while the evaluation of solutions is trivial. This is not the case in CFA-based item selection, because the estimation of CFAs to determine model fit is most likely the most time-consuming part of any iteration.

A number of approaches have adapted ACO algorithms to solve MKP. Leguizamon and Michalewicz (1999) formulated a straightforward adaptation of the AS approach, while Alaya, Solnon, and Ghédira (2004) and Fidanova (2007) propose adaptations of the $\mathcal{MM}$AS. The latter - called ACO-AR - differs from classical $\mathcal{MM}$AS by introducing additional reinforcement to components that were not previously selected making them more likely to be selected in the future. The aim of this approach is to enhance exploration by not punishing choices that were not made in the same way as choices that led to bad solutions. Kong, Tian, and Kao (2008) proposed the Binary Ant System, which is based on the hyper-cube framework proposed by Blum, Roli, and Dorigo (2001). Ke, Feng, Ren, and Wei (2010) proposed another extension of the $\mathcal{MM}$AS which includes dynamic changes of the lower pheromone limit throughout the search to enhance exploration. Recently, Hamann (2015) proposed another $\mathcal{MM}$AS variant called NormANTS, which standardizes pheromones to ensure less problem dependence in the search.

CHAPTER 2

The stuart Approach

This chapter will introduce the stuart approach to item selection and scale shortening. As shown in Chapter 1, current approaches to item selection have some limitations, which this approach strives to address by understanding CFA-based item selection as a I-dimensional multiple knapsack problem with assignment restrictions (IMKAR, see Section 1.3.2) and attempting to solve this problem via an Ant Colony Optimization algorithm based on the $\mathcal{MM}$AS approach (see Section 1.4.3).

This chapter will begin with the problem representation of item selection in the stuart approach. The next section will then give a detailed description of the algorithmic approach employed - introducing two different variants of pheromone localization to the $\mathcal{MM}$AS variant used. This will be followed by a detailed portrayal of item selection in the most simple situation before providing extensions to situations incorporating (a) multiple groups, (b) multiple measurement occasions, and/or (c) multiple sources of information.

The approach shown in this chapter is implemented in a Package for the statistical computing language R (R Core Team, 2016) available at https://bitbucket.org/martscht/stuart. While the development is ongoing, the current version is appended to this thesis. Installation details are provided by the README file. The current version of this package will be submitted to CRAN concurrent to the publication of this thesis and will be available for installation via the CRAN repositories thereafter. The R environment was chosen because of its wide spread in the scientific community. Despite its slower performance in comparison to many other programming languages, R ensures accessibility of the stuart approach to as many interested researchers as possible.

2.1 Problem Representation

As discussed in Section 1.3.2, the initial pool of items is denoted as the set $\mathcal{C}$. Given a situation in which a scale consists of multiple sub-scales (or facets), the pools of items which assess a facet m are denoted $\mathcal{C}_m$. The projection of items from $\mathcal{C}$ to their respective $\mathcal{C}_m$ is not part of the `stuart` approach, but rather a prerequisite step that must be taken following theoretical assumptions that also guided the item construction.

Given the sets of components, the optimization problem $(\mathcal{S}, f, \Omega)$ must be defined. Because the constraints provided by Ω influence the set of possible solutions $\mathcal{S}$, the three parts determining the optimization problem are discussed in reverse order.

Constraints Ω As discussed in Section 1.3.2, the IMKAR posits three specific constraints. The first, ω_1, is given by

$$a_m \geq \sum_{i=1}^{I} w_{im} x_{im}, \tag{1.10, repeated}$$

stating that the sum of weights w associated with the items i selected for a facet m must not exceed the capacity a of that facet. When constructing a questionnaire which can be answered within a specific amount of time (a_m), the time needed to answer each question can be used as weights (w_{im}). However, in most applications all weights will likely be set to $w_{im} = 1$ so that a_m provides the number of items selected for a facet. It is important to note that, because a_m and w_{im} are part of the constraint ω_1 defining the optimization problem, they must be known prior to item selection. In other words, when using $w_{im} = 1$ for all items, the number of items the final scale should have, must be provided. Constraints $\omega_2 := \mathcal{C}_m^s \subseteq \mathcal{C}_m$ and $\omega_3 := \mathcal{C}_m^s \cap \mathcal{C}_{m'}^s \equiv \emptyset \quad \forall m \neq m'$ are imposed upon the selection procedure in general, and are therefore independent of substantive input. How these constraints are reflected in the algorithmic procedure will be discussed in Section 2.2.

Objective Function f The second part of the optimization problem is the objective function f. Throughout Section 1.3 this was referred to only vaguely - as an abstract function of any form, somehow projecting the quality of a solution s onto the set of non-negative real numbers $\mathbb{R}^{\geq 0}$. As stated in Equation (1.11), the classical objective function in IMKAR relates to the benefits of each single component:

$$f(s) = \sum_{m=1}^{M} \sum_{i=1}^{I} b_{ims} x_{im}. \tag{1.11, repeated}$$

In the case of CFA-based item selection, however, the single benefits b_{ims} cannot be known without knowing all possible solutions. This is due to the fact that these benefits are item and solution specific and thus represent benefits only in relation to all possible replacements. However, the sum $\sum_{i=1}^{I} b_{ims}x_{im}$ can be evaluated, because it is simply the quality of the entire solution.

In the case of AS (Section 1.4.2) f was defined as:

$$f(s) = \frac{c}{ll_s}, \qquad \text{(part of 1.18)}$$

where c is some arbitrary constant and ll_s is the total length of the path a solution constructed. The relation to path lengths is due to the fact that these approaches relate to the TSP and other graph problems, but this definition of f can easily be related to the problem of item selection. Because ll_s indicates the *total* path length, it is more akin to $\sum_{i=1}^{I} b_{ims}x_{im}$ than to the specific b_{ims} - much like the results of a CFA, it relates to the entire constructed scale and only to a lesser extent to the specific items.

When using modern approaches to item selection most criteria used to define “good” selections relate either to reliability, validity, or the fit of the measurement model. How these criteria are set in relation to quantify the quality of a solution is mostly dependent upon the objectives a scale is constructed to achieve. When information regarding predictive validity is available, it may be part of f - e.g. in the form of including distal outcomes and optimizing for the R^2 in the prediction of these outcomes by the latent variables underlying the selected items (e.g. Leite et al., 2008). If other scales measuring other constructs are available, it might be useful to include the inverse of the absolute correlation between latent variables of the different scales to maximize discriminant validity. In cases in which a long scale is reduced to a short scale, it may be desirable to reproduce the relationships between the facets in the long version. Janssen et al. (2015) used the correlation between two facets as part of the objective function to reproduce the relationships found in the long version.

Because the `stuart` approach is based on CFA, the most prominent components of f are likely to be approximate indicators of model fit (e.g. RMSEA, SRMR, CFI, TLI) as well as model-based reliability estimates. The former are preferable over tests of model fit (e.g. the χ^2-test of model fit) because they do not include inferential tests. Due to the ACO components of the `stuart` approach, tests concerning model fit should not be performed, because error-rates are likely to be so inflated as to make inference impractical. Instead, descriptive indicators of model fit are preferable. However, most of these fit indexes are not interpreted in a linear fashion - most often being compared to some cut-off criterion derived from simulation studies. Including logistic transformations of fit indexes, instead of the fit indexes themselves, allows for high discrimination at a predefined point, as well as nonlinear benefits of fit. In general, the logistic function is given by

$$g(x) = \frac{u}{1 + e^{-v(x-w)}}, \qquad (2.1)$$

where u is the upper limit of the function, v is the slope of the curve, and w is the x-value of the curve at its maximum discrimination. Therefore, w can be used to influence the point at which one would discriminate most extremely between "good" and "bad" model fit (e.g. at .05 for the RMSEA) and v can be used to influence the strictness of this discrimination. In Section 3.3 the RMSEA is one of the components determining solution quality, with $f(s)$ given by Equation (3.4):

$$g(\text{RMSEA}) = .5 - \frac{.5}{1 + e^{-100(\text{RMSEA} - .05)}}. \tag{2.2}$$

Here, $u = .5$, indicating that at optimal RMSEA its contribution to solution quality is .5, and $w = .05$, indicating that the value of most discrimination is the widespread cut-off value of .05. Due to the large value for v, RSMEA $> .1$ all lead to extremely similar qualities - the same is true for all RMSEA $< .01$. This non-linearity encourages discrimination of solution quality on the basis of other criteria included in f if the model fits "well enough", according to the RMSEA.

The final recommendation with regards to the application specific definition of f concerns models resulting in non-positive definite latent covariance matrices or residual covariance matrices. While the appearance of these cases is sample-specific, it is often indicative of problems that may reappear in other samples. In these cases it is advisable to penalize solutions with regards to their quality. In all applications of the `stuart` approach presented in this thesis, this is done by defining the set of *admissible* solutions $\mathcal{S}^*$. This set contains only those solutions in $\mathcal{S}$ which lead to a converged, proper CFA estimation and positive definite covariance matrices. All solutions not contained in $\mathcal{S}^*$ are automatically assigned $f(s) = 0$, irrespective of the remaining definition of the objective function.

Possible Solutions $\mathcal{S}$ With these rough guidelines for the specific definition of f, the last remaining part of the optimization problem is the set of possible solutions $\mathcal{S}$. This set contains all possible item combinations, which adhere to the constraints provided by Ω and constitute the search space, which is searched using ACO techniques. Under the assumptions that

$$\mathcal{C}_m \cap \mathcal{C}_{m'} \equiv \emptyset \qquad \forall m \neq m', \tag{2.3}$$

stating that no item indicates more than facet, and

$$w_{im} = 1 \qquad \forall i, m, \tag{2.4}$$

indicating that the weights of all items are equal to 1, the number of solutions in $\mathcal{S}$ is given by

$$S = \prod_{m=1}^{M} \binom{I_m}{a_m}, \tag{2.5}$$

where m denotes the facet, M denotes the number of facets, I_m denotes the number of items in the item pool for facet m (i.e. the size of $\mathcal{C}_m$) and a_m denotes the number of items selected for facet m.

2.2 Algorithmic Approach

The algorithm employed in the stuart approach is very similar to the $\mathcal{MM}$AS approach described in Section 1.4.3, but has some key peculiarities which are described in this section.

The implementation of the stuart approach uses what Pedemonte, Nesmachnow, and Cancela (2011) call a master-slave approach to parallelization. More specifically, the stuart implementation can utilize either a coarse-grain or a fine-grain parallelization. Within both approaches, the master process manages all relevant global information such as the current pheromones, the current global best solution, the values of ϕ^{max} and ϕ^{min}, and so on. When using a coarse-grain master-slave parallelization, this information is passed to a slave process, which constructs and evaluates a single solution, meaning that one ant k is performed by a slave. In this way, K ants can be performed by K processors in parallel. This is possible because global information is updated only after each iteration t in $\mathcal{MM}$AS (see Algorithms 2 and 3), meaning that the k ants within each iteration require no information from each other. The results of the K ants are then passed to the master process, which determines s^{ib} (for implementations using the iteration best approach) and whether any ant k in iteration t provides a new s^{gb}. Then, pheromones are updated in the master process before the new slave processes are started. This approach has some obvious advantages over serial implementation, because it is bound to be faster. However, it has the limitation of being able to utilize a maximum of K cores, a limitation not shared by more fine grained master-slave approaches or even cellular approaches in the category system of Pedemonte et al. (2011). The fine-grained approach - also usable within stuart - handles ants in a serial fashion and utilizes parallel processing on the level of the solution evaluation - i.e. in the CFA estimation process. Because the CFA estimation is not part of the stuart approach proper, it is largely dependent on the abilities of the software package used in the CFA estimation process. What Pedemonte et al. (2011) call medium-grain master-slave approaches and cellular approaches are not possible in stuart, because the IMKAR conceptualization in Section 1.3.2 states that the value of single components cannot be assessed without knowledge of the entire solution, and both of these approaches require dividing the problem into subproblems at the level of the solution construction. Of the multi-colony approaches, parallel independent runs can be used with the stuart approach.

In addition to parallelization, stuart is somewhat peculiar in the conceptualization of the

Table 2.1: Minimal example of Φ when localizing pheromones to nodes in a situation with two facets, each with 6 items. Pheromones are initialized to some arbitrary large value - here 999.

Φ_1					
c_1	c_2	c_3	c_4	c_5	c_6
999	999	999	999	999	999
Φ_2					
c_7	c_8	c_9	c_{10}	c_{11}	c_{12}
999	999	999	999	999	999

problem. The ACO approaches discussed in Section 1.4 were defined with regards to problems which are represented as graphs, thereby always making it sensible to deposit pheromones on arcs between nodes. However, in KP it is feasible to deposit pheromones either on arcs (Fidanova, 2003, 2007) or on the nodes themselves (Leguizamon & Michalewicz, 1999). When representing item selection as an IMKAR, both approaches are valid so that both are included in the `stuart` approach. Because the localization of pheromones to nodes is the simpler case, it is discussed first.

2.2.1 Localization to Nodes

When localizing pheromones to nodes, the pheromones are assigned to a set of vectors, denoted Φ. This set contains M vectors, each of size I_m. Just as it is the case for the projection of components from the overall item pool $\mathcal{C}$ into facet specific item pools $\mathcal{C}_m$ (see Section 1.3.2), pheromones are allocated in a facet specific manner. This means, that any given item can have up to M different pheromones associated with it. This guarantees that items are not generally more or less likely to be selected when constructing a solution, but are instead more or less likely to be chosen from $\mathcal{C}_m \rightarrow \mathcal{C}_m^s$. The item-facet specific pheromones are denoted ϕ_{im}. Table 2.1 gives a minimal example of a Φ in a case with two facets, each with 6 items.

As was the case in Algorithms 2 and 3 showing $\mathcal{MM}$AS, the `stuart` approach requires seven input parameters: $\mathbf{H}$, α, β, ρ, T, K, and $p_{s^{gb}}$. Their specific influence on the algorithmic behavior is discussed in more detail in Section 3.1, but it is necessary to investigate the structure of $\mathbf{H}$ here.

Heuristic information is provided in a sets ($\mathbf{H}$), which is of the same structure as its pheromone counterparts Φ. Thus, the Φ_m shown in Table 2.1 could also show heuristics providing the same prior information to each item. As is the case with pheromones, heuristics are provided specifically for each item-facet combination - i.e. as η_{im}.

The probability of choosing a specific item is provided by an adaptation of Equation (1.16):

Algorithm 4: Item selection subroutine implemented in the `stuart` approach when localizing pheromones to nodes.

```
Require: $\Phi(t)$, $\mathbf{H}$, $\alpha$, $\beta$
 1 procedure Item Selection
 2     all $x_{im} \leftarrow 0$
 3     for all $m$ in $(1, ..., M)$ do
 4         while constraint $\omega_1$ in (1.10) do
 5             assign $c_i \in \mathcal{C}_m \to \mathcal{C}_m^s$ using (2.6)
 6             $x_{im} \leftarrow 1$
 7             remove $c_i$ from $\Phi$, $\mathbf{H}$
 8         end while
 9         return $\mathcal{C}_m^s$
10     end for
11     combine $\mathcal{C}_m^s$ to $s$
12     return $s$
13 end procedure
Result: $s$
```

$$p(x_{im} = 1|t) = \frac{[\phi_{im}(t)]^\alpha [\eta_{im}]^\beta}{\sum\limits_{i=1}^{I_m} [\phi_{im}(t)]^\alpha [\eta_{im}]^\beta}, \tag{2.6}$$

which describes the probability of choosing item i in facet m at iteration t as a function of its pheromone ϕ_{im} and its heuristic information η_{im}.

With these parts in place, it is possible to define a subroutine for selecting items, as shown in Algorithm 4. The subroutine requires pheromones at iteration t, heuristics, and their respective non-lineary coefficients. The selection procedure initializes all binary selection indicators (x_{im}) to zero in line 2, indicating that no item has been selected. An item is selected with the probability given by Equation (2.6) in line 5, and the selection indicator of the selected item is set to one (line 6). The pheromone and heuristic vectors are updated to no longer include the item just selected. This is done only as long as the constraint ω_1 is fulfilled - i.e. as long as the sum of the weights of the items does not exceed capacity - and iterated over all facets.

After the items are selected it is necessary to determine $f(s)$. As discussed in Section 2.1, determining an appropriate quality function is part of the theoretical work of each single application. However, it is assumed that $f(s)$ refers to the overall quality of the solution and is not item specific beyond the properties of s. In this case it is possible to generally determine the pheromone update, in line with Equations (1.19) and (1.20), as:

$$\Delta\Phi_m(t, t+1) = \boldsymbol{X}_m^{ib} f(s^{ib}) \tag{2.7}$$

or

$$\Delta\Phi_m(t, t+1) = \boldsymbol{X}_m^{gb} f(s^{gb}) \tag{2.8}$$

depending on whether the iteration-best deposit rule or the global-best deposit rule, described in Section 1.4.3, is used. $\boldsymbol{X}_m$ denotes the vector containing the binary values indicating whether an item c_i was selected in solution s in facet m. Thus, it can be used as a logical filter to replace the cases shown in Equations (1.19) and (1.20).

As is the case for $\mathcal{MM}$AS, an evaporation coefficient ρ guarantees an upper limit for the elements in Φ and, using s^{gb} as a placeholder for the optimal solution s^{opt}, the temporary upper limit is given by

$$\phi^{\max} = \frac{f(s^{gb})}{1-\rho}. \tag{1.22, repeated}$$

Additionally, the temporary lower limit - as discussed in Section 1.4.3 - is provided by:

$$\phi^{\min} = \frac{\phi^{\max}(1 - \sqrt[n]{p_{s^{gb}}})}{avg\ \sqrt[n]{p_{s^{gb}}}} \tag{1.23, repeated}$$

With the limits in place, the pheromone at iteration $t+1$ is given by a variation of Equation (1.26) as:

$$\phi_{im}(t+1) = \begin{cases} \phi^{\min}, & \text{if } \rho\phi_{im}(t) + \Delta\phi_{im}(t, t+1) < \phi^{\min} \\ \phi^{\max}, & \text{if } \rho\phi_{im}(t) + \Delta\phi_{im}(t, t+1) > \phi^{\max} \\ \rho\phi_{im}(t) + \Delta\phi_{im}(t, t+1), & \text{else} \end{cases} \tag{2.9}$$

stating that pheromones exceeding $\phi^{\max}$ are set to the maximum, pheromones lower than $\phi^{\min}$ are set to the minimum, and all other pheromones are updated by evaporation of $\phi_{im}(t)$ and then addition of the $\Delta\phi_{im}(t, t+1)$ determined via Equation (2.7) or (2.8), depending on the deposit rule.

This necessitates the definition of a new convergence criterion as an adaptation of the general $\mathcal{MM}$AS convergence criterion provided in Equation (1.27), specifically:

$$\begin{aligned} &\forall t > 1, i' \neq i'' \quad \exists i', i'' \\ &\qquad (\phi_{i'm}(t) = \phi^{\max}) \wedge (\phi_{i''m}(t) = \phi^{\min}) \wedge \\ &\qquad [(\phi_{im}(t) = \phi^{\max}) \vee (\phi_{im}(t) = \phi^{\min})] . \end{aligned} \tag{2.10}$$

In detail, convergence is achieved if, at any iteration $t > 1$, there is at least one pheromone on items which is minimal, at least one pheromone on items which is maximal, and all pheromones

Algorithm 5: $\mathcal{MM}$AS with the iteration-best deposit rule as implemented in the `stuart` approach.

```
Require: $\mathbf{H}$, $\alpha$, $\beta$, $\rho$, $T$, $K$, $p_{s^{gb}}$
procedure $\mathcal{MM}$AS($\mathcal{S}, f, \Omega$)
    set $1 \ll \phi_{im}(0)$
    $f(s^{gb}) \leftarrow 0$
    while $t \leq T$ and (2.11) is not fulfilled do
        while $k \leq K$ do
            run Item Selection in Algorithm 4
            determine $f(s_k)$
        end while
        $s^{ib} \leftarrow \arg\max f(s_k)$
        if $f(s^{ib}) > f(s^{gb})$ then
            $s^{gb} \leftarrow s^{ib}$
            compute $\phi^{\max}$ with (1.22)
            compute $\phi^{\min}$ with (1.23)
        end if
        determine $\Delta\Phi_m(t, t+1)$ with (2.7)
        compute $\Phi_m(t+1)$ with (2.9)
        $t \leftarrow t + 1$
    end while
    return $s^{gb}$
end procedure
Result: $s^{gb}$
```

are either minimal or maximal. While the limitation to $t > 1$ is not strictly necessary, it is stated here to allow for some tolerance regarding the imprecision in estimating $f(s)$. Including tolerance *tol* factor to allow for this imprecision, rephrases the convergence criterion in Equation (2.10) to

$$\begin{aligned} &\forall t > 1, , i' \neq i'' \quad \exists i' \neq i'' \\ &\qquad (\phi_{i'm}(t) \pm \mathit{tol} = \phi^{\max}) \wedge \left(\phi_{i''m}(t) \pm \mathit{tol} = \phi^{\min}\right) \wedge \\ &\qquad \left[(\phi_{im}(t) \pm \mathit{tol} = \phi^{\max}) \vee \left(\phi_{im}(t) \pm \mathit{tol} = \phi^{\min}\right)\right] \end{aligned} \tag{2.11}$$

The side effect of incorporating *tol* is a quicker convergence, because ϕ_{im} are no longer required to be specific values, but may instead lie in an interval. Because of this interval, however, it is possible to achieve “false convergence” during early iterations. This is the case when the difference between $\phi^{\min}$ and $\phi^{\max}$ is smaller than the tolerance. Because $\phi^{\min}$ is a function of $\phi^{\max}$, as shown in Equation (1.23), this is possible only in few scenarios, which require either very small problems (i.e. selecting from a small pool of items) or very low values of $\phi^{\max}$. Because $\phi^{\max}$ itself is computed using $f(s^{gb})$, this is most likely in very early stages of the algorithm, when only bad solutions have been found.

Algorithm 5 shows the pseudo-code of the algorithm implemented in the `stuart` approach when using the *ib* deposit rule. All pheromones are initialized to an arbitrary large value (line 2) and the initial global-best solution quality is set to be zero (line 3). Then, while neither of the abort criteria is met, the Item Selection procedure is run and the resulting $f(s_k)$ is determined (lines 6 and 7) for K ants in iteration t. The best of these K solutions is stored as the iteration-best solution (line 9). If an s^{ib} is better than the current s^{gb}, it takes its place (line 11) and the pheromone limits are renewed (lines 12 and 13). Finally, the pheromones are updated (lines 15 and 16) and the iteration counter t is increased.

For to the minimal example above, Table 2.1 shows $\Phi(0)$. Assume a simple example in which three of the six items are selected per facet (i.e. $w_{im} = 1$, $a_m = 3$) and $\alpha = \beta = 1$, $\rho = .8$, $T = 5$, $K = 1$, $p_{s^{gb}} = .5$, and no heuristic information is provided. Using the initial pheromones to determine selection probability via Equation (2.6), each item has the same initial probability of being selected: $p(x_{im} = 1|t = 0) = 0.167$. After Item Selection the chosen items are c_1, c_3, c_5, c_7, c_9, and c_{11}. This results in $f(s) = 1$, which is the new s^{ib} and the new s^{gb}. Therefore, the new limits are computed as $\phi^{\max} = 5$ and $\phi^{\min} = 0.175$. $\Delta\Phi(0, 1)$ is shown in Table 2.2. Using Equation (2.9), however, all elements in $\Phi(1)$ are set to 5, because they are limited from above limited by $\phi^{\max}$. In the second iteration, Item Selection chooses the same six items - by chance. Because $K = 1$ this is automatically the s^{ib}. However, because its quality does not exceed the quality of the current s^{gb}, lines 11 through 13 are skipped. Because this is the same solution constructed in the first iteration, $\Delta\Phi(1, 2) = \Delta\Phi(0, 1)$, which is conveniently already depicted in Table 2.2. Using Equation (2.9), $\Phi(2)$ is updated to the version shown in Table 2.2. Because the pheromones of the items that were not selected are multiplied with $\rho = .8$, they decrease from 5 to 4. The pheromones of the items that were selected are also multiplied by ρ, but $\Delta\phi_{im}(1, 2) = 1$, making their pheromones 5 at the end of iteration 2. Because the pheromones are not equal for all items at $t = 2$, their selection probabilities also differ. For the first item selected in Item Selection those with $\phi_{im}(2) = 5$ are chosen with $p(x_{im} = 1|t = 2) = 0.185$, while those with $\phi_{im}(2) = 4$ are chosen with $p(x_{im} = 1|t = 2) = 0.148$. The difference in selection probability then increases as t increases, before the algorithm reaches the convergence criterion or t exceeds T.

Tweaking Algorithm 5 to incorporate the global-best deposit rule results in Algorithm 6. The only notable difference between the two algorithms is the absence of iteration-best solutions in the latter. Instead, all K solutions in iteration t are compared to the current s^{gb} (line 9), which directly inherits s_k if it is better than the current s^{gb}. Irrespective of the deposit rule, a single ant is contained in lines 6 and 7 of the algorithms. Within the **while** loop containing these ants (lines 5 through 8 of both algorithms), no information is updated which is required by the ants - none of the parameters required by the Item Selection process change and the rule with which $f(s)$ is determined, is also fixed. As discussed above, this allows for the parallelization of this portion of the algorithm in the coarse-grain master-slave approach. The fine-grain master slave approach described above is implemented during the evaluation of s (line 7). In both, Algorithm

Table 2.2: Minimal example for a run of the `stuart` approach.

$\Delta\Phi(0,1)$, $\Delta\Phi(1,2)$					
c_1	c_2	c_3	c_4	c_5	c_6
1	0	1	0	1	0
c_7	c_8	c_9	c_{10}	c_{11}	c_{12}
1	0	1	0	1	0

$\Phi(1)$					
c_1	c_2	c_3	c_4	c_5	c_6
5	5	5	5	5	5
c_7	c_8	c_9	c_{10}	c_{11}	c_{12}
5	5	5	5	5	5

$\Phi(2)$					
c_1	c_2	c_3	c_4	c_5	c_6
5	4	5	4	5	4
c_7	c_8	c_9	c_{10}	c_{11}	c_{12}
5	4	5	4	5	4

Algorithm 6: $\mathcal{MM}$AS with the global-best deposit rule as implemented in the `stuart` approach.

```
Require: $\mathbf{H}$, $\alpha$, $\beta$, $\rho$, $T$, $K$, $p_{s^{gb}}$
procedure $\mathcal{MM}$AS($\mathcal{S}, f, \Omega$)
    set $1 \ll \phi_{im}(0)$
    $f(s^{gb}) \leftarrow 0$
    while $t \leq T$ and (2.11) is not fulfilled do
        while $k \leq K$ do
            run Item Selection in Algorithm 4
            determine $f(s_k)$
        end while
        if $f(s_k) > f(s^{gb})$ then
            $s^{gb} \leftarrow s_k$
            compute $\phi^{\max}$ with (1.22)
            compute $\phi^{\min}$ with (1.23)
        end if
        determine $\Delta\Phi_m(t, t+1)$ with (2.8)
        compute $\Phi_m(t+1)$ with (2.9)
        $t \leftarrow t + 1$
    end while
    return $s^{gb}$
end procedure
Result: $s^{gb}$
```

Table 2.3: Minimal example of Φ when localizing pheromones to arcs in a situation with two facets, each with 6 items. Pheromones are initialized to some arbitrary large value - here 999.

			Φ_1			
	c_1	c_2	c_3	c_4	c_5	c_6
c_1	0	999	999	999	999	999
c_2	999	0	999	999	999	999
c_3	999	999	0	999	999	999
c_4	999	999	999	0	999	999
c_5	999	999	999	999	0	999
c_6	999	999	999	999	999	0
			Φ_2			
	c_7	c_8	c_9	c_{10}	c_{11}	c_{12}
c_7	0	999	999	999	999	999
c_8	999	0	999	999	999	999
c_9	999	999	0	999	999	999
c_{10}	999	999	999	0	999	999
c_{11}	999	999	999	999	0	999
c_{12}	999	999	999	999	999	0

5 and Algorithm 6, line 8 requires information about all solutions in iteration t. In theory it is possible to parallelize the determination of $\Delta\Phi_m(t, t+1)$ and the update of the pheromones, because these are facet specific. Because these actions are extremely quick and small, however, this is not included in the current `stuart` implementation.

2.2.2 Localization to Arcs

Localizing pheromones to arcs is a bit more difficult than localizing them to nodes, because it requires extending the conceptualization of the IMKAR beyond the selection independence stated in Equation (1.4) for classical KP, to a version in which the previous selection is memorized and able to influence the probability of the selection immediately after it. This is done by defining a set Φ containing M square matrices. These matrices contain pheromones influencing the choice probability $p(x_{(i,i')m} = 1)$ and denote i in rows and i' in columns, meaning they depict the pheromone for choosing item i' after having chosen item i. In theory, M can be of A_m-dimensionality to encode combination specific choice probabilities (i.e. $p[x_{(i,i',i'',\ldots)m} = 1]$). However, in the `stuart` approach the dimensionality is limited to two, imposing the assumption of conditional independence after controlling for $\sum_{i\in s} w_i x_i$ and the preceding c_i assigned to $\mathcal{C}^s_m$.

Table 2.3 gives an example of Φ when localizing pheromones to arcs for two facets and six items per facet. These matrices are symmetric with zeros along the diagonal. The former reflects the equality of solutions which differ only in the order of items, as stated in Equation (1.3), the latter the fact that no item may be chosen more than once.

Table 2.4: Examples of heuristic information for item combinations in the `stuart` approach with pheromone localization on arcs.

Example 1

	c_1	c_2	c_3	c_4	c_5	c_6
c_1	0	0	0	1	1	1
c_2	0	0	0	1	1	1
c_3	0	0	0	1	1	1
c_4	1	1	1	0	0	0
c_5	1	1	1	0	0	0
c_6	1	1	1	0	0	0

Example 2

	c_1	c_2	c_3	c_4	c_5	c_6
c_1	0	1	3	1	1	1
c_2	1	0	1	1	1	1
c_3	3	1	0	1	1	1
c_4	1	1	1	0	1	3
c_5	1	1	1	1	0	1
c_6	1	1	1	3	1	0

As is the case when localizing pheromones to nodes, heuristics also have the same structure as the pheromones when localizing to arcs. Note that, while heuristics and pheromones have the same structure, the assumptions of symmetry and zero-diagonals are not imposed on $\mathbf{H}$. Specific heuristics are denoted $\eta_{(i,i')m}$ and have the possibility to relate to combinations of items. This is one of the central aspects when choosing a localization, because it determines the degree in which heuristic information can be provided. If the item selection should be limited because of properties of item combinations, the arc localization is necessary. This might be the case when some items in the original item pool are so similar, that at most one of them should be included in the final solution, or when items are phrased positively and negatively and the final solution should consist of as many positively as negatively phrased items. Example 1 in Table 2.4 shows a situation in which items 1, 2, and 3 are positively phrased, items 4, 5, and 6 are negatively phrased and provides the heuristic matrix necessary to ensure an even number of positive and negative items. While filling $\mathbf{H}$ with zeros and ones provides logical gates to item selection, it might also be necessary to make selection of an item more likely when a different one was previously selected. This is done in Example 2 of Table 2.4, where items 1 and 3 as well as 4 and 6 are 3^β-times more likely to be chosen together.

Due to this structure of Φ and $\mathbf{H}$ the probability of selecting an item i' directly after having selected an item i is given by an adaptation of Equation (1.16):

Algorithm 7: Item selection subroutine in `stuart` when localizing pheromones to arcs.

```
Require: $\Phi(t)$, $\mathbf{H}$, $\alpha$, $\beta$
 1 procedure ITEM SELECTION
 2     all $x_{(i,i')m} \leftarrow 0$
 3     for all $m$ in $(1, ..., M)$ do
 4         randomly assign a starting item $c_i \in \mathcal{C}_m \rightarrow \mathcal{C}_m^s$
 5         $x_{im} \leftarrow 1$
 6         while constraint $\omega_1$ in (1.10) do
 7             choose $i'$ from $i$ using (2.12)
 8             assign $c_{i'} \in \mathcal{C}_m \rightarrow \mathcal{C}_m^s$
 9             $x_{(i,i')m} \leftarrow 1$
10             determine $x_{i'm}$ with (2.13)
11             remove $i'$ columns from $\Phi$, $\mathbf{H}$
12         end while
13         return $\mathcal{C}_m^s$
14     end for
15     combine $\mathcal{C}_m^s$ to $s$
16     return $s$
17 end procedure
Result: $s$
```

$$p(x_{(i,i')m} = 1|t) = \frac{[\phi_{(i,i')m}(t)]^\alpha [\eta_{(i,i')m}]^\beta}{\sum\limits_{i'=1}^{I_m} [\phi_{(i,i')m}(t)]^\alpha [\eta_{(i,i')m}]^\beta}. \tag{2.12}$$

With the selection probability in place, the item selection subroutine is shown in Algorithm 7. It should be noted that the selection of items is coded in $x_{(i,i')m}$ - i.e. on the arcs - but the constraint in ω_1 pertains to the items themselves. This can be included by defining $x_{i'm}$ as the marginal sum of $x_{(i,i')m}$ for all i. Or, more concisely:

$$x_{i'm} = \sum_{i}^{I} x_{(i,i')m}. \tag{2.13}$$

As was the case when localizing pheromones to nodes, the item selection procedure begins with the initialization of the selection indication to zero. Within each facet a random item is chosen as the starting location (line 4) and its selection indicator is set to 1 (line 5). Then, while the sum of weights does not exceed capacity, a path is chosen from i to some i' with the probability determined via Equation (2.12), and the component $c_{i'}$ is assigned to the facet specific component set of solution s (lines 7 and 8). The arc between i and i' is coded as chosen (line 9) and the marginal sum is calculated (line 10), to make the evaluation of ω_1 possible. In the final step, all columns in $\mathbf{H}$ and Φ pertaining to the selected item i' are removed (line 11).

After the selection of items, each solution is evaluated and the quality $f(s)$ is determined. As was the case for localizing pheromones to nodes, the pheromone updates are computed via:

$$\Delta\Phi_m(t,t+1) = \boldsymbol{X}_m^{ib} f(s^{ib}) \tag{2.7, repeated}$$

or

$$\Delta\Phi_m(t,t+1) = \boldsymbol{X}_m^{gb} f(s^{gb}) \tag{2.8, repeated}$$

depending on the pheromone deposit that is utilized. With the pheromone update in place, the pheromones at iteration $t+1$ are then determined as:

$$\phi_{(i,i')m}(t+1) = \begin{cases} \phi^{\min}, & \text{if } \rho\phi_{(i,i')m}(t) + \Delta\phi_{(i,i')m}(t,t+1) < \phi^{\min} \\ \phi^{\max}, & \text{if } \rho\phi_{(i,i')m}(t) + \Delta\phi_{(i,i')m}(t,t+1) > \phi^{\max} \\ \rho\phi_{(i,i')m}(t) + \Delta\phi_{(i,i')m}(t,t+1), & \text{else} \end{cases} \tag{2.14}$$

whereby the pheromones on the selection of i' after selecting i are $\phi^{\min} \leq \phi_{(i,i')m} \leq \phi^{\max}$. With these limits the convergence criterion is defined as:

$$\begin{aligned} &\forall t > 1, i \neq i', i \neq i'' \neq i''' \quad \exists i'', i''' \\ &\qquad \left(\phi_{(i,i'')m}(t) \pm \mathit{tol} = \phi^{\max}\right) \wedge \left(\phi_{(i,i''')m}(t) \pm \mathit{tol} = \phi^{\min}\right) \wedge \\ &\qquad \left[\left(\phi_{(i,i')m}(t) \pm \mathit{tol} = \phi^{\max}\right) \vee \left(\phi_{(i,i')m}(t) \pm \mathit{tol} = \phi^{\min}\right)\right]. \end{aligned} \tag{2.15}$$

stating that for every item i there is at least one i'', for which the pheromone on the arc is maximal, at least one item i''' for which it is minimal, and all the pheromones leading to any i' are either maximal or minimal. As was the case for localization to nodes, the tolerance *tol* is included to accommodate imprecision in the estimation of $f(s)$.

Algorithms 8 and 9 depict the algorithmic procedure when localizing pheromones to arcs in the stuart approach using either the iteration-best or the global-best deposit rule. The algorithms themselves do not differ from those shown in Algorithms 5 and 6 for the localization of pheromones to nodes, but their pieces differ somewhat.

The most notable differences in the localization approaches are handled in the Item Selection subroutine shown in Algorithm 7. The other two differences are the convergence criteria, referenced line 4 of both algorithms, as well as the computation of pheromones at $t+1$, referenced in lines 16 and 15, respectively.

Algorithm 8: $\mathcal{MM}$AS with the iteration-best deposit rule as implemented in the `stuart` approach.

```
Require: $\mathbf{H}$, $\alpha$, $\beta$, $\rho$, $T$, $K$, $p_{s^{gb}}$
procedure $\mathcal{MM}$AS($\mathcal{S}, f, \Omega$)
    set $1 \ll \phi_{im}(0)$
    $f(s^{gb}) \leftarrow 0$
    while $t \leq T$ and (2.15) is not fulfilled do
        while $k \leq K$ do
            run Item Selection in Algorithm 7
            determine $f(s_k)$
        end while
        $s^{ib} \leftarrow \arg\max f(s_k)$
        if $f(s^{ib}) > f(s^{gb})$ then
            $s^{gb} \leftarrow s^{ib}$
            compute $\phi^{\max}$ with (1.22)
            compute $\phi^{\min}$ with (1.23)
        end if
        determine $\Delta\Phi_m(t, t+1)$ with (2.7)
        compute $\Phi_m(t+1)$ with (2.14)
        $t \leftarrow t + 1$
    end while
    return $s^{gb}$
end procedure
Result: $s^{gb}$
```

Algorithm 9: $\mathcal{MM}$AS with the global-best deposit rule as implemented in the `stuart` approach.

```
Require: $\mathbf{H}$, $\alpha$, $\beta$, $\rho$, $T$, $K$, $p_{s^{gb}}$
procedure $\mathcal{MM}$AS($\mathcal{S}, f, \Omega$)
    set $1 \ll \phi_{im}(0)$
    $f(s^{gb}) \leftarrow 0$
    while $t \leq T$ and (2.15) is not fulfilled do
        while $k \leq K$ do
            run Item Selection in Algorithm 7
            determine $f(s_k)$
        end while
        if $f(s_k) > f(s^{gb})$ then
            $s^{gb} \leftarrow s_k$
            compute $\phi^{\max}$ with (1.22)
            compute $\phi^{\min}$ with (1.23)
        end if
        determine $\Delta\Phi_m(t, t+1)$ with (2.7)
        compute $\Phi_m(t+1)$ with (2.14)
        $t \leftarrow t + 1$
    end while
    return $s^{gb}$
end procedure
Result: $s^{gb}$
```

2.3 Measurement Models

The representation of the problem of item selection as well as the algorithmic approach presented in the previous two sections can be applied to any approach to evaluating s and determining a meaningful $f(s)$. Within the `stuart` approach, classical test theory (CTT; e.g. Lord & Novick, 1968; Novick, 1966; Steyer, 1989) and CFA are used to determine $f(s)$ by establishing measurement models and evaluating their results.

The basics of CFA are discussed in a multitude of introductory textbooks (e.g. Brown, 2015; Kline, 2011; Schumacker & Lomax, 2016). In the `stuart` approach the flexibility of CFA is utilized to allow for item selection in situations with (*a*) multiple facets, (*b*) multiple groups, (*c*) multiple occasions, and (*d*) multiple sources of information, or any combination of the four. The general advantage of including this flexibility in the process of item selection is that questionnaires can be constructed with the objective of measurement invariance across groups explicitly included, for example. This flexibility, in combination with the understanding of item selection as a combinatorial problem (see Section 1.3), allows for a straightforward manner of selecting items in complex situations without the need for multi-stage decision processes of the test constructor, thereby overcoming some of the major problems of modern CFA driven item selection.

The current implementation of the `stuart` approach utilizes either lavaan (Rosseel, 2012) or Mplus (L. K. Muthén & Muthén, 1998-2015) to estimate the underlying CFA models. While both are extremely similar in their implementation of standard estimation approaches to structural equation modeling, they do differ in some degrees that may have an impact on the results of item selection using the `stuart` approach. First, they differ in their possibilities regarding parallelization. As discussed in Section 2.2, it is possible to use either a coarse-grain or a fine-grain master-slave parallelization approach. Which of these is used is dependent upon the estimation software, with the coarse-grain approach being used with lavaan and the fine-grain approach being used with Mplus. This is due to the fact, that lavaan does not provide parallel computing capabilities within its estimation process as of version 0.5-22, thus making the coarse-grain approach necessary in this instance. By contrast, using Mplus in the solution evaluation requires the start of an external process when using R, thereby necessitating the storage of an Mplus input file, an Mplus output file, and a data file for each single ant k. As of writing, a coarse-grain parallelization of this process was slower than the fine-grain parallelization implemented im Mplus Version 7.3. In addition to parallelization, the two software packages differ in which fit criteria they provide, thereby influencing the possible definitions of the objective function f. It should be noted that, irrespective of the estimation software, in the current implementation of `stuart` standard errors are not computed for any solutions during the process of item selection. This is done because it enhances the speed of the algorithm immensely.

The rest of this section will give an overview of the `stuart` approach to including information from the four constellations listed above. The most simple form, selecting items to multiple facets, will be discussed first and then extended to include the other three conditions. Note, that this presentation is limited to continuous indicators, but can be extended to situations with categorical indicators. This extension would then allow for the inclusion of IRT models via their SEM counterparts without necessitating an explicit reformulation of the fundamentals of this thesis.

2.3.1 Item Selection for Simple Situations

In this section a “simple situation” is understood as a situation in which items are selected for a scale, when considering only a single occasion, a single group, and a single source of information. The inclusion of multiple facets is inherent to the representation of the problem of item selection in the `stuart` approach (see Section 2.1). Specifically, because the problem of item selection is construed as an IMKAR (as discussed in Section 1.3.2), multiple facets are a basic part of the optimization problem.

In the most general setting, the selected items are modeled as

$$c_{oim} = \tau_{im} + \lambda_{im}\xi_{om} + \epsilon_{oim}, \tag{2.16}$$

where c_{oim} represents observation o of the component (i.e. the item) i of facet m, τ_{im} represents the intercept, λ_{im} represents the factor loading on the latent variable ξ_{om}, representing the underlying construct of facet m, and ϵ_{oim} is the measurement error.

Equivalently, Equation (2.16) can be formulated in matrix notation as

$$\boldsymbol{C} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\epsilon} \tag{2.17}$$

with the model implying the covariance structure

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}^\top + \boldsymbol{\Theta}, \tag{2.18}$$

and the meanstructure

$$\boldsymbol{\mu} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\kappa} \tag{2.19}$$

where $\boldsymbol{\Lambda}$ is the matrix of factor loadings, $\boldsymbol{\Psi}$ is the latent covariance matrix, $\boldsymbol{\Theta}$ is the residual covariance matrix, $\boldsymbol{\tau}$ is the vector of of intercepts, $\boldsymbol{\mu}$ is the vector of item means, and $\boldsymbol{\kappa}$ is the vector of latent means.

Within the `stuart` approach strict unidimensionality is assumed. As described in Section 2.1 and stated in Equation (1.13), one of the constraints imposed onto the IMKAR is that no item can be selected to indicate more than one facet. In the strict interpretation utilized in the `stuart` approach, this means that

$$\lambda_{im'} = 0 \qquad\qquad \forall c_{im} \in \mathcal{C}^s_m, m \neq m', \tag{2.20}$$

or stated a different way, that all elements in $\mathbf{\Lambda}$ are zero if they represent cross-loadings of indicators on latent variables that are not the facet they were selected for. With the additional assumption of uncorrelated errors, writeable as

$$cov(\epsilon_{oim}, \epsilon_{oi'm'}) = 0 \qquad\qquad \forall (i,m) \neq (i', m'), \tag{2.21}$$

which states that $\mathbf{\Theta}$ is a diagonal matrix, the measurement model is overidentified for most cases with either four components in $\mathcal{C}^s$ and one facet or two components in each $\mathcal{C}^s_m$, when the latent variables ξ_{om} are correlated (e.g. Brown, 2015). These guidelines do not guarantee identification of the measurement models however, as they remain susceptible to empirical underidentification (e.g. when factor loadings or latent covariances are zero). In situations with overidentified models, the quality function f can contain indexes of model fit. When the model is just identified - e.g. when three items are selected for one facet - other criteria can be included in the quality function, but model fit will always be perfect, and will not differentiate between different solutions, making the inclusion of fit-criteria unnecessary.

It should be noted, that the assumptions stated in Equations (2.20) and (2.21) are not strictly necessary when constructing tests, but should be striven for. Specifically in the `stuart` approach, the inclusion of correlated residuals and cross-loadings is hard to justify, given that their extent is dependent on the specific combination of items. If items are assumed to be correlated above and beyond the strict unidimensionality due to substantive considerations, it might be better to consider either including separate facets for these or extending the approach to a case with multiple sources of information. For example, a case in which some items are negatively phrased, while others are positively phrased, may be more suited to a model in which a positive and a negative facet are defined as different sources of information for the same latent construct using an MTMM approach. This extension is discussed in more detail in Section 2.3.4.

Five different measurement models are considered in the `stuart` approach for simple situations, each implementing different invariance restrictions within a facet m (cf. Eid, Gollwitzer, & Schmitt, 2015; Steyer, 1989; Steyer & Eid, 2001).

The most general model is that which is most often assumed in applications of CFA: the model of τ-congeneric variables. In this case, all items within a facet are simply related to the latent variable by Equation (2.16), implying that item easiness (indicated by τ_{im}), item discrimination (indicated by λ_{im}), as well as item reliability (indicated by $1 - \frac{var(\epsilon_{oim})}{var(c_{oim})}$), may differ across all items. Constraining the factor loadings to be equal for all items in the same facet ($\lambda_{im} = \lambda_{i'm}$) results in a model of essentially τ-equivalent measures. In this case items within the same facet discriminate equally on the latent dimension, indicating that an individual's score on an item differs from that on a different item of the same facet only by a constant and measurement error. Additionally constraining $\tau_{im} = \tau_{i'm}$ results in a model of τ-equivalent measures, which indicates identical item easiness for all items of the same facet and that an individual's values on two items of the same facet differ only due to measurement error. Restraining $var(\epsilon_{im}) = var(\epsilon_{i'm})$ without the assumption of invariant intercepts, results in a model of essentially τ-parallel measures. The implications of this model are that all items pertaining to facet m have the same reliability and that - much like for the model of essentially τ-equivalent measures - differences between items are some constant (namely $\tau_{im} - \tau_{i'm}$) and measurement error. In the case of τ-parallel measures, residual variances and intercepts are equal, meaning that item scores differ only due to measurement error and that all items have the same reliability.

Within the `stuart` approach all five models are applicable to any facet. This means, that the assumptions of a given measurement model apply to a facet m as a whole, but may differ for different facets of the same questionnaire. One major difference between the `stuart` approach and most approaches to CFA is that a measurement model is assumed a priori. In classical CFA the different measurement models are often tested sequentially using likelihood-ratio tests or information criteria. Using this procedure, the most restrictive measurement model which is not significantly worse than a less restrictive model is assumed as best. In the `stuart` approach this is not the case. Instead, a measurement model is assumed a priori and items are selected to maximize the objective function given this measurement model. This way, if a scale is not in line with the assumed measurement model, this is visible in overall model fit indexes, which would result in lower $f(s)$ when they are included in the objective function.

It should be emphasized that the `stuart` approach aims at finding the best solution under the circumstances provided. The measurement model is one of these circumstances - an application specific constraint which must be provided - thus putting the "search" for a good measurement model beyond its scope. This should not be interpreted as a shortcoming because it is intentional. Within the `stuart` approach it is necessary to construct a hypothetical, ideal scale - formulating all of its properties - and then searching for a solution in the pool of item combinations which best represents this ideal.

2.3.2 Multiple Groups

As pointed out in the previous section, using the `stuart` approach to select items from multiple facets is ingrained in the IMKAR representation of item selection. Extending this approach to include multiple groups is straightforward, because the process of item selection itself is not influenced by the inclusion of multiple groups. Instead it is assumed that:

$$\mathcal{C}_g = \mathcal{C}_{g'} \tag{2.22}$$

and

$$(\mathcal{S}_g, f_g, \Omega_g) = (\mathcal{S}_{g'}, f_{g'}, \Omega_{g'}), \tag{2.23}$$

meaning that the set of components $\mathcal{C}$ and the optimization problem $(\mathcal{S}, f, \Omega)$ are independent of the grouping variable. Under this assumption, the optimization problem is still the one described in Section 2.1 and the additional assumptions made in Equations (2.20) and (2.21) also apply to the case of multiple groups.

However, the inclusion of multiple groups necessitates the extension of Equation (2.16) to multiple-group CFA (e.g. Jöreskog, 1971; Muthén, 1989) by including group-specific parameters:

$$c_{oimg} = \tau_{img} + \lambda_{img}\xi_{omg} + \epsilon_{oimg}, \tag{2.24}$$

where the new indicator g denotes the group, making all model parameters group dependent. For multiple groups this implies the covariance structure

$$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Psi}_g \boldsymbol{\Lambda}'_g + \boldsymbol{\Theta}_g, \tag{2.25}$$

and the meanstructure

$$\boldsymbol{\mu}_g = \boldsymbol{\tau}_g + \boldsymbol{\Lambda}_g \boldsymbol{\kappa}_g, \tag{2.26}$$

meaning that the covariance matrices and mean vectors of all groups g are independent of each other, because the observations are assumed to stem from independent groups.

Measurement invariance in the assessment of individuals from multiple populations is one of the central research areas in psychological methodology and psychometrics. While the core concepts of measurement invariance in factor analysis are not new (e.g. Drasgow & Kanfer, 1985; Meredith, 1964, 1993), research on appropriateness and implementation of measurement invariance is still ongoing (e.g Millsap, 2011; Raykov, Marcoulides, & Li, 2012; van de Schoot, Schmidt, & De Beuckelaer, 2015, for an overview of the current developments) and some of these newer approaches aim at relaxing invariance restrictions by either adopting partial measurement invariance schemes (Steinmetz, 2013) or utilizing approximate invariance (van de Schoot et al., 2013). Because the `stuart` approach is aimed at constructing scales, these less restrictive approaches are not directly utilized and a more conservative approach is implemented, based on the invariance levels described by Meredith (1993).

In line with Meredith (1993), the `stuart` approach includes four invariance levels for the item selection in multiple groups.

The first, configural invariance, assumes that zero-value restrictions on factor loadings λ_{img} are the same across all values of g, or more formally:

$$\lambda_{im'g} = \lambda_{im'g'} = 0 \tag{2.27}$$

for all components selected for facet m with $m \neq m'$. This is a straightforward combination of the assumptions stated in Equation (2.20), (2.22), and (2.23) within the `stuart` approach. This indicates that the structure of measurements is equivalent across groups, but the parameters themselves may vary. While this may seem very unrestrictive, the establishment of configural invariance is a point of contention in many areas of cross-cultural research (e.g. van de Vijver, van Hermert, & Poortinga, 2008). The second level is weak factorial measurement invariance, given by

$$\lambda_{img} = \lambda_{img'} = \lambda_{im} \tag{2.28}$$

stating that the factor loadings are independent of the grouping variable g. This means that the relationships between latent variables may be compared across groups (Widaman & Reise, 1997), because the relationship between a latent variable and its indicators is the same across groups, and that differences between individuals in the same group are comparable across groups. Because of the equality of factor loadings, items discriminate to an equal degree on the latent constructs across different groups. Strong factorial invariance, the third form of invariance, is achieved if the equality of the intercepts

$$\tau_{img} = \tau_{img'} = \tau_{im} \tag{2.29}$$

is included in addition to the equality of loadings stated in Equation (2.28). This allows for the estimation of $G - 1$ means of latent variables, which implies

$$\mu_{img} - \mu_{img'} = \lambda_{im}(\kappa_{mg} - \kappa_{mg'}). \tag{2.30}$$

This states that the intercepts of all manifest variables are independent of g and that the means of the observed variables differ only by the scaled differences in latent variable means. These restrictions allow for the comparison of latent means between groups. Additionally, items have the same discrimination on the latent variable and the same difficulty in different groups. The final level of invariance implemented in the `stuart` approach is strict factorial invariance, which is given in a model with strong factorial invariance and

$$var(\epsilon_{oimg}) = var(\epsilon_{oimg'}) = var(\epsilon_{oim}) \tag{2.31}$$

as an additional assumption. The invariance assumption placed on the residual variances means that all group differences in the manifest variables are due solely to differences in the latent variables and their relations. All differences in the manifest variance-covariance matrix are due to group differences in latent variances and covariances and all differences in the means of items are due to differences in the latent means. However, it should be noted that even strict factorial invariance does not imply full measurement invariance (Millsap, 2011, p. 106). Reliabilities may be different across groups, because they depend on the distribution of the latent variables.

As was the case for measurement models described in Section 2.3.1, the invariance assumptions are commonly tested sequentially, stopping at the most restrictive, tenable invariance assumption (see Vandenberg & Lance, 2000, for a review of classical approaches). Again, this is not the case here. As stated above, the `stuart` approach requires the statement of an ideal model - if a scale is intended for use across different populations this must be reflected in the assumptions about the solution and integrated into the process of item selection. With an assumption about measurement invariance across groups in place, the `stuart` approach can be used to search for a solution meeting those requirements. Again, it should be emphasized that these assumptions can be provided on a facet-by-facet basis, if there is reason to believe that some facets are likely less invariant in their measurements. As pointed out at the beginning of this section, configural invariance is assumed as soon as this approach is applied to data with multiple groups.

2.3.3 Multiple Occasions

Despite the intuitive approach of understanding the incorporation of longitudinal measurement settings into the process of item selection as an extension of the `stuart` approach, it actually represents a restriction of the basic approach discussed in Section 2.3.1. This becomes clearer when looking at the measurement invariance restrictions associated with CFA for repeated measurements.

Measurement invariance assumptions made in longitudinal modeling are described in the same four steps as those stated in Section 2.3.2 for multiple groups, but do not pertain to samples from different populations, but instead to different variables representing repeated measures of each other (T. D. Little, 2013). As was the case for multiple groups, the first invariance assumption states configural invariance, meaning that the structure of factor loadings is the same across multiple occasions. This implies that the same items must be selected to repetitions of the same measurement when assuming configural invariance.

In an application of the `stuart` approach with repeated measures and no configural invariance, the relation of variables as repeated measures is not discernible from the relation of any other two variables, and the item selection would proceed as described in Section 2.3.1. In this case repeated measures allowing for the selection of different components simply constitute different sets $\mathcal{C}_m$, each with their own allocation from $\mathcal{C}_m \rightarrow \mathcal{C}_m^s$.

To introduce configural invariance it is necessary to constrain the selection $\mathcal{C}_m \rightarrow \mathcal{C}_m^s$ and $\mathcal{C}_{m'} \rightarrow \mathcal{C}_{m'}^s$ for those facets m and m' which are believed to be repeated measures. This is done by introducing the set $\mathcal{R}$ which contains the sets of repeated measures $\{m, m', ...\}$ for which the item selection is constrained to be the same. Each set of repeated measures is denoted $\mathcal{R}_v$ and the constraint

$$\boldsymbol{X}_m = \boldsymbol{X}_{m'} \qquad \forall m, m' \in \mathcal{R}_v \tag{2.32}$$

is introduced, meaning that the selection matrix $\boldsymbol{X}$ (as introduced in Section 2.2.1) is identical for all facets m and m' which are defined as repeated measures in the same $\mathcal{R}_v$. This equality of selection implies that selection probabilities as stated in Equation (2.6) for localizing pheromones to nodes and in Equation (2.12) for localizing pheromones to arcs of repeated measures must be the same. This means that:

$$p(x_{im} = 1|t) = p(x_{im'} = 1|t) \qquad \forall m, m' \in \mathcal{R}_v \tag{2.33}$$

$$p(x_{(i,i')m} = 1|t) = p(x_{(i,i')m'} = 1|t) \qquad \forall m, m' \in \mathcal{R}_v. \tag{2.34}$$

Because the selection probability is proportional to the product $\Phi(t)^\alpha \mathbf{H}^\beta$, its equality also indirectly implies the equality of pheromones and heuristics for the same items from repeated measures.

Note, that this introduction of $\mathcal{R}$ means that, even in cases without repeated measures, all m are assigned to a set $\mathcal{R}_v$. In these cases $V = M$, because each facet is the first occasion of measurement for its dimension and each m is assigned to its own $\mathcal{R}_v$, so as not to impose any restrictions.

In addition to this constraint in the item selection, it is necessary to assume that all zero-value restrictions are the same across measurement occasions to constitute a model fulfilling configural measurement invariance.

It should be noted that the constraint stated in Equation (2.32) allows for flexible definitions of repeated measures, because no restrictions are imposed on the $\mathcal{R}_v$ themselves. Thereby, it is possible to apply this constraint in different fashion to different facets, making it possible to select items in situations in which some facets are measured more often than others (e.g. for the simultaneous item selection for a trait and a state scale) or situations in which configural invariance is assumed for some facets, but not for others.

As was the case for item selection in multiple groups, the remaining invariance assumptions concern only the CFA models and not the actual process of item selection. In line with T. D. Little (2013) weak factorial invariance is given if

$$\lambda_{im} = \lambda_{im'} \qquad \forall m, m' \in \mathcal{R}_v, \tag{2.35}$$

meaning that the factor loadings of all items are equal for repeated measures. As was the case for multiple groups, this implies the same item discrimination across multiple occasions and allows for the comparison of the relationships between latent variables across occasions.

Strong factorial invariance assumes the equality of intercepts and freedom of $V-1$ latent means:

$$\tau_{im} = \tau_{im'} \qquad \forall m, m' \in \mathcal{R}_v \tag{2.36}$$

implying

$$\mu_{im} - \mu_{im'} = \lambda_{im}(\kappa_m - \kappa_{m'}) \qquad \forall m, m' \in \mathcal{R}_v, \tag{2.37}$$

which indicates that the difficulties of items are the same over different occasions but that the means of indicators may change due to changes in the latent means, making the detection of latent mean change possible. Finally, strict factorial invariance supposes that

$$var(\epsilon_{oim}) = var(\epsilon_{oim'}) \qquad \forall m, m' \in \mathcal{R}_v, \tag{2.38}$$

meaning that error variances are equal for repeated measures, indicating that changes in the distribution of items over time are solely due to changes in the latent variables.

As was the case in Section 2.3.2, it is necessary to impose a level of measurement invariance for all facets prior to item selection. As pointed out above, if configural invariance should not be assumed, two repeated facets can simply be defined as two regular facets without including them in the same $\mathcal{R}_v$. Beyond the assumption of measurement invariance, including repeated measures into item selection has the positive effect of allowing for new components in the quality function f. Specifically, it allows to go beyond the optimization of the reliability in state-assessment and instead optimize for the reliability in change assessment between occasions, or to maximize the change sensitivity of solutions.

2.3.4 Multiple Sources of Information

The incorporation of multiple sources of information into the `stuart` approach is very similar to that of including multiple occasions. However, it differs in the implementation of the absence of measurement invariance. In the case of repeated measures, assuming no invariance results in the simple case described in Section 2.3.1. Two facets m and m' are simply handled as two completely different facets, because they are indistinguishable from those cases in terms of CFA. However, the approach used to accommodate multiple sources of information in this thesis has implications for the model, even when the item selection process is not constrained in any fashion, thereby allowing for different item pools for different assessment methods.

The `stuart` approach utilizes a restricted version of the CTC(M-1) model for multitrait-multimethod (MTMM) data (Eid, 2000; Geiser, Eid, & Nussbeck, 2008). The CTC(M-1) approach uses the declaration of a reference method and the contrast of all other (non-reference) methods against this reference to model MTMM data. This means that a state factor is defined as the reference-measured state and residual method factors represent the common aspects of items stemming from non-reference sources of information that are not shared with the reference method (Eid, 2000). In the restricted version presented by Geiser et al. (2008) this is achieved by regressing the latent states of non-reference methods on the reference-measured states, thereby creating latent residual variables. This approach to modeling multiple sources of information is, however, limited to structurally different methods. The implications of this and possible extensions beyond this are discussed in more detail in Section 5.3.

Within the `stuart` approach, this restricted CTC(M-1) approach is represented by

$$\xi_{om} = \beta_{0m} + \beta_{1m}\xi_{om'} + \zeta_{om} \tag{2.39}$$

where ξ_{om} is the observation o of the latent variable for facet m, $\xi_{om'}$ is that for the reference source m', β_{1m} is the regression weight predicting ξ_{om} by $\xi_{om'}$, β_{0m} is the intercept, and ζ_{om} is the observation on the residual method factor of facet m. Note that the identifiability of β_{0m} is subject to the identification approach used for the mean structure in a model. When the mean structure is identified by constraining all latent means to 0, β_{0m} must too be 0 if no additional restrictions are imposed on the mean structure (as is the case in configural and weak measurement invariance).

This regression allows for the determination of consistency coefficients between the reference method m' and each m, conceived of as a non-reference method measuring the same construct. Additionally, the computation of the inverse - the method specificity - is also possible. Both of these coefficients can be used in the objective function f to incorporate aspects of the agreement between different sources of information, which is the sole reason for the restriction imposed unto the latent covariance matrix in this setting. If the consistency between different sources of information assessing the same construct is not of interest to item selection (i.e. if it is not part of the objective function), the restrictions imposed by the CTC(M-1) approach presented here are not necessary, and different methods can be handled in the unrestricted fashion as different facets

m and m', for which the latent correlations are freely estimated. In fact, the two approaches - the inclusion of the restricted CTC(M-1) approach vs. the free estimation of latent correlations - differ only if additional facets and accompanying constraints are included in the measurement model.

Configural invariance between different sources is included via the definition of a set $\mathcal{B}$, containing sequences of $(m_u^*, m, ...)$ facets which are multimethod assessments of the same latent constructs, denoted $\mathcal{B}_u$. Note, that while for repeated measures $\mathcal{R}$ contains sets, $\mathcal{B}$ contains sequences - meaning that the order of elements is irrelevant in cases with multiple occasions, while it is of relevance for multiple sources of information. This is due to the CTC(M-1) approach setting a reference method, as indicated above. The first element in each $\mathcal{B}_u$ is denoted m_u^*. As is the case for single occasions, having only one source of information (or no invariance between sources of information) does not prevent subsets m from being assigned to sets $\mathcal{B}_u$. It simply means that each m is the first and only element m_u^* in its own $\mathcal{B}_u$ and that $M = U$.

The configural invariance assumption is imposed via a straightforward transfer of Equation (2.32) to the sets of MTMM facets:

$$\boldsymbol{X}_m = \boldsymbol{X}_{m'} \qquad \forall m, m' \in \mathcal{B}_u, \tag{2.40}$$

whereby the same items must be chosen for all facets indicated as being different sources of the same latent construct in a shared $\mathcal{B}_u$. As was the case for longitudinal applications, this implies identity of the choice probabilities and - to a lesser extent - the equality of pheromones and heuristics for these facets.

The remaining three invariance assumptions are very rare in applications with multiple sources of information, but can be incorporated analogously to those stated in Section 2.3.3. For weak factorial invariance this means

$$\lambda_{im} = \lambda_{im'} \qquad \forall m, m' \in \mathcal{B}_u. \tag{2.41}$$

Strong factorial invariance is given if

$$\tau_{im} = \tau_{im'} \qquad \forall m, m' \in \mathcal{B}_u, \tag{2.42}$$

which results in the identifiability of β_{0m} in Equation (2.39) in cases in which the mean structure is identified via $\kappa_m = 0$, making

$$\mu_{im} = \lambda_{im}\beta_{0m}. \tag{2.43}$$

Strict factorial invariance is given if

$$var(\epsilon_{oim}) = var(\epsilon_{oim'}) \qquad\qquad \forall m, m' \in \mathcal{B}_u. \tag{2.44}$$

While the factorial invariance assumptions are rarely of relevance, they can be incorporated if need be. The main advantage of explicitly incorporating the MTMM structure into the `stuart` approach, instead of using the simple approach (Section 2.3.1) for cases without configural invariance, and the approach for multiple occasions (Section 2.3.3) for all others, is the explicit incorporation of the consistency coefficients between sources of information into the objective function f. However, due to the asymmetric nature of the restricted CTC(M-1) approach employed here, it is necessary to select a reference method a priori - in addition to choosing an invariance level.

2.3.5 The Full Approach

The four previous sections each presented the `stuart` approach in specific situations. This final section is dedicated to combining these situations to a full approach, allowing for any combination of multiple facets, multiple groups, multiple occasions, and multiple sources of information.

Consider the following minimal, fictional example: a scale is constructed for the assessment of two, closely related constructs - e.g. cognitive and affective empathy in adolescents. Because it is known that empathy can be somewhat gender-specific, the scale needs to ensure that measurements are invariant across the two populations - allowing differences between genders only in the latent variables. Additionally, because there is considerable interest in the development of these latent constructs during adolescence, it is crucial that the scale be measurement invariant over time. Because of this, two measurement occasions are used during the process of scale construction - one in freshman year and one in sophomore year of high school. Finally, it is of interest to generate a self- and a parent-report version of the scale to investigate the consistency between the empathy judged by the self and the empathy detected by another. The initial item pool consists of 37 items (17 for cognitive and 20 for affective empathy) and the final scale is intended to consist of four items per facet. In this minimal example, there are eight facets $m = (1, ..., 8)$, each indicating a specific construct-occasion-method combination. Table 2.5 shows the facet designation in the fourth column, labeled m. Without any further restrictions - meaning without assuming configural invariance for either repeated measures or multiple sources - Equation (2.5) can be used to determine the number of possible solutions as 1.77×10^{28}.

Table 2.5: Facet designation m and allocation to sets of repeated measures $\mathcal{R}_v$ and sources $\mathcal{B}_u$, as well as to the partition $\mathcal{Q}_h$ in the minimal example.

Construct	Occasion	Method	m	$\mathcal{R}_v$	$\mathcal{B}_u$	$\mathcal{Q}_h$
Affective	Freshman	Self	1	$\mathcal{R}_1 = \{1{,}3\}$	$\mathcal{B}_1 = (1{,}2)$	$\mathcal{Q}_1 = \{1{,}2{,}3{,}4\}$
		Parent	2	$\mathcal{R}_2 = \{2{,}4\}$	$\mathcal{B}_1 = (1{,}2)$	$\mathcal{Q}_1 = \{1{,}2{,}3{,}4\}$
	Sophomore	Self	3	$\mathcal{R}_1 = \{1{,}3\}$	$\mathcal{B}_2 = (3{,}4)$	$\mathcal{Q}_1 = \{1{,}2{,}3{,}4\}$
		Parent	4	$\mathcal{R}_2 = \{2{,}4\}$	$\mathcal{B}_2 = (3{,}4)$	$\mathcal{Q}_1 = \{1{,}2{,}3{,}4\}$
Cognitive	Freshman	Self	5	$\mathcal{R}_3 = \{5{,}7\}$	$\mathcal{B}_3 = (5{,}6)$	$\mathcal{Q}_2 = \{5{,}6{,}7{,}8\}$
		Parent	6	$\mathcal{R}_4 = \{6{,}8\}$	$\mathcal{B}_3 = (5{,}6)$	$\mathcal{Q}_2 = \{5{,}6{,}7{,}8\}$
	Sophomore	Self	7	$\mathcal{R}_3 = \{5{,}7\}$	$\mathcal{B}_4 = (7{,}8)$	$\mathcal{Q}_2 = \{5{,}6{,}7{,}8\}$
		Parent	8	$\mathcal{R}_4 = \{6{,}8\}$	$\mathcal{B}_4 = (7{,}8)$	$\mathcal{Q}_2 = \{5{,}6{,}7{,}8\}$

First, integrating the multiple group capabilities into the approach for simple situations, results in the measurement equation shown in Equation (2.27). While this indicates that the CFA models are group specific, the process of item selection is not different from the one used for simple situations, as shown in Section 2.2. Specifically, because of the assumptions stated in Equations (2.22) and (2.23), items are selected independently of the grouping variable - enforcing the same selection in both groups. The multiple group aspect is introduced only to the analysis of the constructed solution, i.e. the CFA conducted. Thereby, the `stuart` approach assumes that the same items measuring affective and cognitive empathy need to be selected in the current example. To ensure measurement invariance, strong factorial invariance could be imposed, for example, allowing for differences in residual variances and latent means between the groups.

Because the combination of multiple measurement occasions and multiple groups do not interfere in any way, both can be included into the approach independently. The restrictions resulting from some facets m and m' being repeated measures of each other are primarily imposed on the selection process by Equation (2.32). This means, that for each specific combination of construct and source, the same items must be selected across the occasions. By imposing this constraint (i.e. configural invariance over time), repeated measures of the constructs are assigned to their respective sets $\mathcal{R}_v$ as shown in the fifth column of Table 2.5 - labeled $\mathcal{R}_v$. Due to the imposed equality of item selection for facets in the same set of repeated measures, the problem size decreases. Instead of being a function of the number of facets M it is now a function of the number of sets of repeated measures V, updating Equation (2.5) to:

$$S = \prod_{v=1}^{V} \binom{I_m}{a_m} \tag{2.45}$$

where the binomial coefficient is computed for one (and only one) $m \in \mathcal{R}_v$ for each $\mathcal{R}_v$. This results in 1.33×10^{14} possible solutions in this example. Imposing more restrictive measurement invariance is part of the CFA and does not change the process behind item selection. Instead, it is noticeable only in the quality of solutions $f(s)$, if the objective function f contains an indicator of model fit. This change in quality influences *which* solutions are constructed - via its influence on item selection probabilities as stated in Equations (2.7) and (2.8) - but not *how* solutions are constructed. The imposition of configural invariance, on the other hand, changes *how* solutions are constructed as described above and in Section 2.3.3.

Extending the example to include multiple sources of information (self- and parent-reports) is similar to the process of including multiple occasions. As described in Section 2.3.4, facets that stem from different sources of information but measure the same construct are allocated to sets $\mathcal{B}_u$ if they are at least configurally invariant. When this is combined with multiple measurement occasions, only facets measuring the same construct at the same occasion are assigned to a common $\mathcal{B}_u$ and item selection is restrained via Equation (2.40). This is shown in column six of Table 2.5.

Using only multiple sources (and ignoring repeated measures, for now), reduces the number of possible combinations in the same way that the use of multiple occasions does:

$$S = \prod_{u=1}^{U} \binom{I_m}{a_m} \tag{2.46}$$

which results in the same number of possible combinations for the example, because there are as many sources of information as there are measurement occasions. However, combining both approaches reduces the problem size even more, because the number of unique facets for which it is necessary to select items is only two, in this case. Table 2.5 shows that, via the constraints imposed on repeated measures, $m = 1$ and $m = 3$ will use the same item selection matrix $\boldsymbol{X}_m$. Additionally, $m = 2$ is a different source of information for $m = 1$, whereby these facets share the same selection matrix. Consequently, only one full run of Item Selection in either Algorithm 4 or Algorithm 7 is necessary for each construct, because the remaining three facets are known once one is given. Figure 2.1 illustrates the facet allocation for this example.

In general the number of possible solutions is a function of the number of partitions on the combined set of $\{\mathcal{R}_v, \mathcal{R}_{v'}, ..., \mathcal{R}_V, \mathcal{B}_u, \mathcal{B}_{u'}, ..., \mathcal{B}_U\}$. This means the number of subsets for which

$$\mathcal{Q}_h \cap \mathcal{Q}_{h'} = \emptyset \qquad \forall h \neq h' \tag{2.47}$$

every pair is disjoint, and

$$m \in \mathcal{Q}_h \tag{2.48}$$

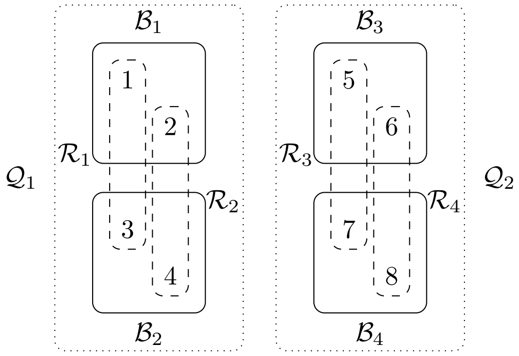


Figure 2.1: Illustration of facet allocation with eight facets, as per the example shown in Table 2.5. Sets of repeated measures $\mathcal{R}_v$ are depicted with dashed lines, sets of different sources of information $\mathcal{B}_u$ are depicted with solid lines, and the partitions $\mathcal{Q}_h$ are depicted with dotted lines.

all m are element of a partition $\mathcal{Q}_h$. Note that while this is defined here for the combination of multiple occasions and multiple sources of information, it is also true for all cases that are simplifications - i.e. all of the cases presented. This is because the case with only a single source of information simply assigns each facet m to its own set $\mathcal{B}_u$, such that $U = M$. In this example this would mean that there would be four partitions in $\mathcal{Q}$, because there are four sets in $\mathcal{R}$ which do not overlap. Therefore, in the most general case, the number of possible solutions in the `stuart` approach is given by

$$S = \prod_{h=1}^{H} \binom{I_m}{a_m}. \tag{2.49}$$

In the example shown in Table 2.5 this amounts to 11531100 possible solutions. While this example is nicely symmetrical, note that this need not be the case. If, for example, the parent-report for cognitive empathy were only assessed for freshmen the last two lines of Table 2.5 would simply be removed and the lower two lines would read $\{5\}$ and $\{6\}$ in the column $\mathcal{R}_v$ to indicate that the parent-reports of cognitive empathy measured at the freshman age are repeated measures of no other facets.

To summarize the full approach, recall the description of item selection as a MKP in Section 1.3.2. That section described how items are assigned from the overall item pool $\mathcal{C}$ to facet specific item pools $\mathcal{C}_m$ and items are then chosen from those facet specific item pools to the facet solutions $\mathcal{C}^s_m$ (see Figure 1.2). In this section, constraints on the selection from $\mathcal{C}_m \to \mathcal{C}^s_m$ were imposed by defining relations between facets m. This was done by assigning each facet m to a set of facets $\mathcal{R}_v$ consisting of those m which are repeated measures of each other. Simultaneously each facet m is assigned to a set of facets $\mathcal{B}_u$ containing those facets which are multiple sources of the same information. The sets $\mathcal{Q}_h$ were then defined as partitions of the combined set of $\mathcal{R}$ and $\mathcal{B}$, which contain facets for which the same items must be selected. Figure 2.1 shows this facet allocation diagrammatically. With these sets in place, the algorithmic procedure described in Section 2.2 is performed - and items are only selected for one of the facets in each $\mathcal{Q}_h$ and copied to all other facets in the same $\mathcal{Q}_h$, due to the constraints stated in Equations (2.32) and (2.40).

CHAPTER 3

Parameter Evaluation

The purpose of this chapter is twofold: (a) evaluate the overall performance of the `stuart` approach to item selection and scale shortening proposed in this thesis, and (b) derive recommendations concerning parameter settings for applications of this approach. To achieve these two goals an evaluation study utilizing a wide array of parameter constellations is performed for the shortening a scale with a relatively complex internal structure in a simple data constellation.

As described in Section 2.3.1, simple data structures constitute situations with only one occasion, one group, and one assessment method. Because the evaluation presented here relies on replications of a number of different parameters, such a data constellation was chosen to make results obtainable in a reasonable time frame. The relatively complex internal structure is provided by the Ryff-Scale, which consists of six specific facets. More detail on this scale is given in Section 3.2.

Because the aim of this study is not just the investigation of the performance of the approach, but also the derivation of guidelines for parameter settings in applications, the evaluation is done in a two-step procedure. First, a number of constellations of constant parameter settings, derived from the literature investigating the performance of ACO algorithms in other situations, is evaluated. In the second step the results of this evaluation are used to derive parameter schedules to be evaluated. The possible combinations of parameter schedules are simply too many for an exhaustive study, therefore only those that seem promising, according to the results of the evaluation of constant parameter settings, are investigated more thoroughly.

This chapter will begin with an overview of the parameters and their expected influence on the performance of the general $\mathcal{MM}$AS algorithm before giving an introduction into the scale under investigation. This is followed by a description of the performance measures used to evaluate the behavior of the `stuart` approach. Afterwards, the constant parameter settings are described in detail and the results from these situations are presented. These results are then used in the elicitation of parameter schedules to be investigated in the second phase of the evaluation. After a presentation of the results regarding parameter schedules, the chapter will conclude with a brief overall discussion and recommendations regarding parameter settings for applications of the `stuart` approach.

3.1 Tuneable Parameters

As shown in Algorithms 5 through 9, as well as throughout Section 2.2, the $\mathcal{MM}$AS algorithm used in this approach requires a set of seven parameters as well as the definition of the pheromone function $f(s)$ to work properly. This section will provide a description of these parameters and their expected influence on the behavior of the algorithm itself. Additionally, the final section will provide an overview of parameter schedules which are relevant to the second half of this evaluation.

3.1.1 Heuristic Information H

The heuristic information provided in **H** is a determining factor in the construction of the solution, because it guides the algorithm in its search. As shown in Section 2.2, the selection probability of each item is determined via

$$p(x_{im} = 1|t) = \frac{[\phi_{im}(t)]^{\alpha}[\eta_{im}]^{\beta}}{\sum\limits_{i=1}^{I_m}[\phi_{im}(t)]^{\alpha}[\eta_{im}]^{\beta}}, \tag{2.6, repeated}$$

when localizing pheromones to nodes. Localizing pheromones to arcs gives a very similar equation for determining selection probabilities - see Equation (2.12) - but this evaluation focuses on a case with node localization. Both equations clearly indicate, that heuristic information is directly linked to selection probability and that this link is independent of any solution's quality. Instead, heuristic information is a constant attractiveness of a choice (be it the node or the arc), biasing the search in a certain direction. In classical applications of ACO algorithms to routing problems (such as the TSP), the heuristic information generally includes information about the route length or cost, thereby being directly related to the quantity that is being optimized for (e.g. Dorigo & Stützle, 2004; Dorigo & Stützle, 2010; Pellegrini, Favaretto, & Moretti, 2006; Stützle, 1998). For cases solving MKP, heuristic information is often related to cost and benefit of a component (e.g. Leguizamon & Michalewicz, 1999; Fidanova, 2007). This information is often chosen, because it is readily available and seems sensible for guiding the search towards good solutions, since it is directly related to the problem and the objective function.

In the case of item selection, it can prove a bit more difficult to derive meaningful heuristics, however. As pointed out in Section 1.3, the quality of a solution can only be derived from the full selection - partial solutions provide relatively little insurance about the success of the final solution. While a routing problem can provide useful information after each single step, i.e. the distance traveled can be divided into partial solutions, this is not the case for item selection. While a selection of four items may provide very convincing fit, adding a fifth item can ruin the fit of the measurement model or invalidate it completely (e.g. due to Heywood-Cases). This implies, in reverse, that providing heuristics derived from partial solutions may steer the search in the wrong direction, leading to scenarios where heuristic information is actually detrimental to the search process. This contradicts the way heuristics are used in traditional ACO applications, where the use of heuristics is sometimes necessary (e.g. when using the classic AS approach; Dorigo & Stützle, 2004) and always recommended. This does not mean, that heuristics are generally useless for item selection. The objective function can contain other criteria, which may be indicated by uni- or bivariate information about the items. Facet specific item-total correlations can provided heuristic information about item reliability, correlations of items with external criteria can be used when the objective function contains correlations between the latent variables and some distal outcome, or item inter-correlations can be used to indicate the direction to more homogeneous selections. Beyond this, as discussed in Chapter 2 heuristic information can be also be of dichotomous nature, filtering variables to make specific combinations impossible or defining specific anchor items. As pointed out in Section 2.2, the heuristic information that is included in item selection can determine the localization of pheromones. Using only item-specific information makes node localization feasible, while using information about the combinations of items makes it necessary to use arc localization.

Because of the comparatively unclear relation between heuristics and solution quality in the `stuart` approach, determining to effect of heuristic information on the selection process is difficult. Providing heuristic information that is not in line with the best solution, from the perspective of the objective function, will likely lead to worse solutions. However, the coefficients α and β (discussed in the following two sections) allow fine grained control over the balance between statistical item quality, assessed via the quality function, and a priori item quality provided by the heuristics in directing the search.

3.1.2 Non-Linearity of Pheromones α

As shown in Equations (2.6) and (2.12), the parameter α determines the extent with which pheromones influence choice probabilities. Because of this, higher values of α are accompanied by less exploration.

Assume a facet has six items and three are selected, with the pheromone being limited to the range $[0, 1]$. Figure 3.1 shows the influence of α on the selection probability of an item, if all other items have a deposited pheromone of .5. Higher values of α lead to substantially increased selection probability of items that have higher ϕ_{im}. Specifically, if all $\eta_{im} = 1$, then

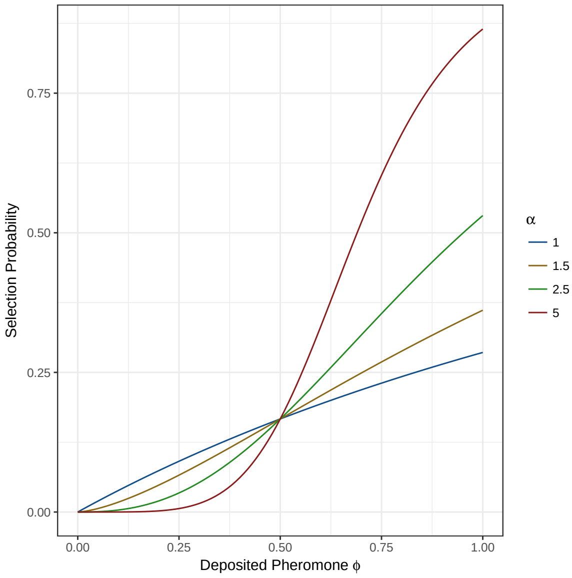


Figure 3.1: Selection probability of an item dependent on deposited pheromone ϕ_{im} for different values of α.

the odds-ratio of choosing item i and not item i' is given via Equation (2.6) as

$$\left[\frac{\phi_{im}(t)}{\phi_{i'm}(t)}\right]^{\alpha}, \tag{3.1}$$

making α the non-linearity coefficient of this selection relation. Because of this, items are selected more consistently with larger values of α, limiting exploration while enforcing exploitation by forcing many solutions to contain high-value items.

In general, $\alpha > 0$ is used in applications of any ACO algorithm to ensure that higher pheromones lead to higher selection probability. In many cases, large values of α are discouraged, because they have been found to lead to too fast convergence (Dorigo & Stützle, 2004), often ending in recommendations of $\alpha = 1$ for most applications (Alaya et al., 2004; Favaretto, Moretti, & Pellegrini, 2009; Stützle, 1998; Stützle et al., 2010; Wei, 2014). However, it has also been proposed, that the value of α is less relevant to the explorative behavior of $\mathcal{MM}$AS, and focus should lie on the ratio of $\frac{\beta}{\alpha}$ instead (Pellegrini et al., 2006).

3.1.3 Non-Linearity of Heuristic Information β

As can be seen in Equations (2.6) and (2.12), the effect of β on the heuristics is akin to that of α on the pheromones. In general, higher values of β incentivize the selection of heuristically favored items or combinations of items. As Pellegrini et al. (2006) point out, the actual value of β is less relevant than the relation between α and β, because the latter determines how to weigh information against each other. When pheromones and heuristics lead to conflicting results the ratio of α to β determines which information is more valuable in constructing solutions.

As mentioned in Section 3.1.1, the relevance of heuristics is not as clear cut in item selection as in many other instances. Because the information provided by $\mathbf{H}$ may contradict the favorable choices in terms of Φ after a number of iterations, β might be considered a prime target for parameter adaptation. By decreasing β throughout the search, heuristics can guide the search towards areas that are likely to contain very good solutions, while the pheromones can then be used to search within those areas. It should be noted, that the specific value of $\beta > 0$ is irrelevant if $\mathbf{H}$ is used as a logical filter, meaning that it implies impossible combinations (by allocating zeros in the localization to arcs) or anchor items (by allocating very large values to some items). In turn, any thoughts about the setting of β are only of relevance when relative heuristics are provided, making some selections somewhat more likely than others.

Thus, previous studies on the influence of β in other areas of application are only mildly informative. For many applications a setting of $\beta = 2$ is recommended for $\mathcal{MM}$AS (Pellegrini, Stützle, & Birattari, 2010; Stützle, 1998), though values of up to $\beta = 5$ have also been proposed (Escario et al., 2015; Wei, 2014), as have recommendations of $\beta = 1$ (Favaretto et al., 2009). In general, most recommendations fall between $2 \leq \beta \leq 5$ (Dorigo & Stützle, 2004).

However, in most previous applications the heuristic information (and its weight, in the form of β) had been derived from the same information used to determine the solution quality in f. Under these circumstances, the objective of β is to ensure faster convergence to optimal or near-optimal solutions. In the `stuart` approach, heuristics and β may be difficult to obtain, as described above, increasing the possibility of providing sub-optimal heuristics, i.e. heuristics biasing the search in a way that is not actually better than providing no prior information.

3.1.4 Evaporation Coefficient ρ

While the parameters discussed in the previous section are directly related to choice probabilities, the evaporation coefficient ρ is more directly related to the behavior of the algorithm iteself. Recall from Section 2.2.1, that the update rule for pheromones is given by

$$\phi_{im}(t+1) = \begin{cases} \phi^{\min}, & \text{if } \rho\phi_{im}(t) + \Delta\phi_{im}(t,t+1) < \phi^{\min} \\ \phi^{\max}, & \text{if } \rho\phi_{im}(t) + \Delta\phi_{im}(t,t+1) > \phi^{\max} \\ \rho\phi_{im}(t) + \Delta\phi_{im}(t,t+1), & \text{else} \end{cases} \tag{2.9}$$

when localizing to nodes and very similarly when localizing to arcs, as shown in Equation (2.14). Thus, the coefficient ρ determines the speed with which pheromones of bad choices - or choices that simply have not been made - disappear, making these less likely to be made in the future.

Because the pheromones are initialized to some arbitrary large number and then scaled to $\phi^{\max}$ at the first iteration, and the algorithm converges only if at least one of the pheromones is minimal - see Equations (2.11) and (2.15) - ρ is crucially important in determining runtime and exploration. Large values (e.g. $\rho = .99$ or $\rho = .95$) lead to slow convergence and emphasize exploration, because the pheromone on choices disappears very slowly, while the opposite is true for low values. Specifically, if a choice does not acquire any pheromone during the entire run, it will reach the threshold $\phi^{min} + tol$ required for convergence after a bare minimum of

$$\left\lceil \frac{\log\left(\frac{\phi^{min}+tol}{\phi^{max}}\right)}{\log(\rho)} \right\rceil \tag{3.2}$$

iterations. Therefore, when $\rho = 0$ the runtime is one iteration, because pheromones are not stored and the current pheromones depend only directly on the pheromones of the previous solution. Any choice not made in the first iteration would automatically receive $\phi_{im}(t) = \phi^{\min}$, any choice made would automatically receive $\phi_{im}(t) = \phi^{\max}$ and convergence would be achieved instantly. On the other hand if $\rho = 1$, Equation (1.22), and therefore the $\phi^{\max}$, is not defined.

Regarding any values in between the extremes, different values have been proposed to lead to near-optimal solutions in adequate runtime. In most cases, values of $\rho \geq .9$ are proposed when using no local search (Alaya et al., 2004; Dorigo & Stützle, 2004; Fidanova, 2007; Pellegrini et al., 2010; Stützle et al., 2010) in $\mathcal{MM}$AS strategies, though values as low as $\rho = .7$ have been found to be optimal in some situations (Favaretto et al., 2009). Alternatively, Hamann (2015) proposes using an iteration-sensitive value for ρ and Wei (2014) proposes fine-tuning ρ in increments of .1. Overall, large values ($\rho \geq .95$) have been identified as leading to near-optimal solutions, albeit in long runtimes (Alaya et al., 2004; Dorigo & Stützle, 2004; Stützle et al., 2010).

3.1.5 Number of Iterations T

The number of iterations is used solely to define the abort criterion stated in the Algorithms presented in Section 2.2. Because the `stuart` approach utilizes an $\mathcal{MM}$AS algorithm, the abort criterion is only relevant when convergence is slow.

Slow convergence may be due to a number of factors. As pointed out in the previous section, the evaporation coefficient ρ determines the lower limit of runtime. Using large values of ρ will exponentially increase runtime, because the $\mathcal{MM}$AS approach relies on evaporation of the pheromones to achieve convergence. As pointed out previously, small values of α and β lead to more exploration and thus to longer runtimes. A combination of these parameter settings might make it necessary to abort the search if - after a reasonable number of iterations - no new s^{gb} has been found. What constitutes a "reasonable" number of iterations is largely problem dependent. Larger problems indicate a necessity for larger values of T, because more iterations are necessary to explore the larger problem space. Therefore, T is imposed as a limit defined mainly by the resources (mostly time) available for any given application. Because T can translate to actual runtimes, an important aspect of choosing T is the time required to evaluate the solution, i.e. the time required for the CFA estimation.

Given the probabilistic nature of solution construction, there is no guarantee of convergence for any instance of item selection using the `stuart` approach. Imagine a situation in which two solutions differing only in one item allocation lead to the same $f(s)$. In this instance an optimally behaving run of `stuart` would not converge, because the pheromones allocated to these two items could not both be maximal at same time[1] and it is very unlikely for either to evaporate to $\phi^{\min}$. To avoid infinite construction of solutions with very similar, near-optimal quality, the abort criterion on the number of iterations T is used.

3.1.6 Number of Ants K

The number of ants influences runtime as well as exploration due to the fact that $\mathcal{MM}$AS utilizes an elitist strategy. Because in each iteration t there are K solutions and at most one of these (depending on the deposit rule) is allowed to deposit pheromone, the choice of K can be a difficult balance between runtime and solution quality. In applications to routing problems, it is often recommended to use as many ants as there are nodes to visit via the route (e.g. Dorigo & Stützle, 2004; Escario et al., 2015), while values of as low as one quarter of the number of cities have also been suggested as optimal (Wei, 2014). For MKP no clear recommendation is discernible, with evaluation studies using between 10 and 100 ants per iteration (Alaya et al., 2004; Fidanova, 2007; Hamann, 2015; Ke et al., 2010).

Stützle et al. (2010) point out, that low values tend to perform best during the early phases of the search, while higher values tend to perform better during later phases. This is due to the fact that all ants of the same iteration use the same pheromones as their guide. Thus, in early stages it may be beneficial to establish pheromone trails on many promising solutions by using lower numbers of ants, thereby increasing the proportion of ants depositing pheromones and keeping the number of ants operating with poor pheromones low. In later stages of the search, good pheromone trails are established and more ants in each single iteration can use the same information to explore specific, promising areas of the solution space.

As stated in Section 2.2, the `stuart` approach can utilize a coarse-grain approach to parallelization, meaning that the ants of an iteration are evaluated in parallel. In these cases, it should be considered to utilize a multiple of the number of processing units available to minimize overhead.

[1] Note that this is possible, when setting the tolerance in the convergence criteria, stated in Equations (2.11) and (2.15), to $tol > (1-\rho)\phi^{\max}$.

3.1.7 Probability of the Global Best Solution $p_{s^{gb}}$

The upper and lower limits imposed on pheromones were discussed in Section 1.4.3. The upper limit $\phi^{\max}$ is given by

$$\phi^{\max}(t) = \frac{f(s^{gb})}{1-\rho} \tag{1.22, repeated}$$

and the lower limit is given by

$$\phi^{\min} = \frac{\phi^{\max}(1 - \sqrt[n]{p_{s^{gb}}})}{(avg-1)\sqrt[n]{p_{s^{gb}}}}. \tag{1.23, repeated}$$

Because these limits are of central importance in the convergence criteria given by Equations (2.11) and (2.15), they greatly influence convergence speed and the quality of the final solution. However, the definitions of the limits each contain only one parameter which can be manipulated: ρ and $p_{s^{gb}}$. The influence of ρ is discussed in Section 3.1.4.

Of the limits, $\phi^{\min}$ has been found to have greater influence on convergence speed of $\mathcal{MM}$AS (Stützle, 1998; Stützle & Hoos, 2000), which is mainly due to the initialization of the pheromones to $\phi^{\max}$, necessitating the pheromone decay all the way down to the lower limit. As discussed in Section 1.4.3, n indicates the total number of decisions and avg indicates the average number of choices at each decision. In item selection as described in Chapter 2, $n = \sum_{h=1}^{H} I_m$ for one m per partition $\mathcal{Q}_h$, indicating one unique facet for which I items are selected. The avg is simply given by the mean of the number of available items at each of the selections made. Thus, both of these variables are given by the size of the problem, and are not manipulatable.

The probability of finding the current global-best solution, however, is a free parameter, that must be set for each application. While any choice of arbitrarily small value of $p_{s^{gb}}$ will improve performance over not implementing a lower limit, the actual value that leads to best performance is dependent on the size of the problem at hand (Stützle, 1998). This can be seen quite readily, by imagining two situations. In the first, a single-facet scale is constructed by choosing 4 items from a pool of 10 items. Using Equation (2.5), this results in 210 possible combinations. By pure chance, the s^{gb} will be constructed with $p = 0.005$. In the second situation, imagine the example described in Section 2.3.5, where a two-facet scale was constructed by selecting four out of 17 and 20 items, respectively. In this case, the random chance of constructing s^{gb} is $p = 8.67 \times 10^{-8}$. The $p_{s^{gb}}$ represents the probability of constructing s^{gb} after the algorithm has converged - thus it should be set to a much higher value in the first case, than in the second. Note, that setting $p_{s^{gb}} = 1$ will result in a $\phi^{\min} = 0$, thus ridding the algorithm of the lower pheromone limit and the possibility of convergence, when no tolerance factor is included.

3.1.8 Parameter Schedules

As pointed out for some of the parameters in the previous sections, the desirable effects of parameter settings on the algorithmic procedure of the `stuart` approach may change throughout the course of the search. During early phases it is desirable to search the space efficiently, finding promising areas of the search space and exploring many different possibilities. During later stages of the search, the focus often lies more on exploiting the knowledge generated about the search space in the most efficient manner possible, avoiding the construction of bad solutions by searching only those areas that are most probable to contain good solutions. Therefore, it may be desirable to have different parameter settings early (e.g. using a small K to enhance the communication between ants) than during later stages of the search (e.g. using lower values of ρ to quickly forget bad solutions).

In general, three types of parameter flexibility can be distinguished: (a) fixed parameter settings, which do not change at all, (b) scheduled parameter settings, which follow a pre-defined schedule, and (c) adaptive parameter settings, which change dependent on the current results provided by the algorithm. In line with the classification taxonomy proposed by Eiben, Michalewicz, Schoenauer, and Smith (2007) for evolutionary algorithms, the former two options represent parameter *tuning*, while the third represents parameter *control*. Approaches which tune parameters require prior knowledge about parameters - be it from studies using similar algorithms to investigate similar problems or from repeatedly solving the same problem with different parameter settings - while approaches using parameter control will react to intermediary results of the algorithm, thereby adapting themselves to the problem at hand.

Intuitively, it makes most sense to assume that approaches using adaptive parameter settings should perform best, because many of the parameters detailed above were said to be problem specific. Using approaches which control parameters within the specific application seem most promising from this perspective because, as Pellegrini, Stützle, and Birattari (2012, p. 23) state, "an instance-optimal parameter setting always obtains better results than any other setting". Therefore, using strategies to derive these instance-optimal settings on the fly would be expected to vastly outperform general recommendations for parameter settings. In experimental studies specifically investigating $\mathcal{MM}$AS, however, Pellegrini et al. (2010) found adaptive strategies to be no better than scheduled parameter settings, while Pellegrini et al. (2012) even found them to be detrimental to overall performance in comparison to (informed) fixed parameter settings. This is most likely due to the increase in complexity of the problem which needs to be solved. In cases with adaptive parameter settings, the problem at hand is larger than the actual problem that needs to be solved, because it encompasses the dimensions of parameters (Eiben et al., 2007).

Stützle et al. (2010) compared the performance of fixed and scheduled parameter settings in solving the TSP via $\mathcal{MM}$AS and the ACS, concluding that the use of parameter schedules can lead to substantial decreases in runtime, without substantial losses in the quality of the final solution. It should be noted here, that this means that in this study fixed parameter settings lead to the highest solution quality, albeit at a much slower pace. Additionally, Stützle et al. (2010) found that parameter schedules changing values from "high" to "low" settings at fixed points in the algorithm outperform those where parameters are linear functions of runtime.

Due to the similarity of problems tackled with the `stuart` approach (i.e. they are all IMKAR and all use the same problem representation), this chapter aims at providing recommendations for parameters and parameter schedules which can be used as guidelines in applications.

3.2 Ryff-Scale for Psychological Well Being

The evaluation study is performed on a dataset stemming from an intervention study in the field of positive psychology (Tempel, 2016). The sample consists of 1506, predominantly female (89.83%) participants of the ages between 18 and 79 ($M_{age} = 35.72$, $SD_{age} = 12.06$). Though the intervention study itself is longitudinal in nature, only data from the first measurement occasion (prior to the intervention) are used in this evaluation. For a much more thorough description of the study and the sample please see Tempel (2016).

Among many other scales, the participants answered a German translation of the 54-item version of the Ryff-Scale for the assessment of psychological well being (Ryff, 1989; Risch, Strohmayer, & Stangier, 2005), which is the scale under investigation in this evaluation. This scale was assessed via an online questionnaire on a 6-point scale ranging from "1 = decisively disagree" to "6 = decisively agree". Despite the online assessment, non-repsonse was low on the items of the Ryff-Scale, at just 0.4% of all responses, but quite a sizeable proportion of participants failed to respond on at least one of the 54 items (14.41%).

The Ryff-Scale was chosen here, because it fulfills a number of criteria: (*a*) it is quite long with a total of 54 items, (*b*) it has a complex structure with six theoretically distinguishable facets, (*c*) it has been investigated heavily, and (*d*) there is considerable need for a short version.

Regarding points (*a*) and (*b*), Table 3.1 shows the item allocation to the six facets. The entire scale used by Tempel (2016) is provided in Appendix B. The six facets were initially proposed by Ryff (1989) as theoretical dimensions of psychological well-being and the scale was then constructed in several steps to reflect these theoretical considerations. It is important to note that these facets were not derived empirically from a common pool of items in a single study - as is often the case in psychological questionnaire construction - but rather that the scale was constructed with the explicit goal of assessing these six theoretically distinct dimensions. This does not mean, however, that the differentiation of these facets is unanimously accepted (e.g. Abbott et al., 2006; Kafka & Kozma, 2002; van Dierendonck, 2004; Ryff & Singer, 2006; Springer & Hauser, 2006; Springer, Hauser, & Freese, 2006). In fact, Springer and Hauser (2006) found the facets self-acceptance, purpose in life, personal growth, and environmental mastery to be so highly correlated when correcting for item-phrasing effects using a method factor, that they

Table 3.1: Theoretical factor structure of the 54-item Ryff-Scale. Negatively phrased items are *emphasized.*

Facet	Item Number										No. of Items
Self-Acceptance	4	9	*14*	23	28	*31*	*43*	45	48	51	10
Positive Relations	1	*5*	*10*	15	*24*	*32*	34	*39*	47		9
Autonomy	6	11	*16*	19	*25*	35	*40*	*44*	52		9
Environmental Mastery	2	*7*	*12*	17	*20*	29	36	*49*	53		9
Purpose in Life	*8*	*13*	*22*	*27*	*30*	33	38	42	46		9
Personal Growth	*3*	*18*	21	*26*	37	*41*	*50*	*54*			8

proposed these four dimensions may not be discernible empirically.

While the discussions regarding the empircal internal structure of the Ryff-Scale are still ongoing, this scale has also been applied in a variety of settings and translated to a multitude of different languages (e.g. Burns & Machin, 2009; van Dierendonck, 2004; van Dierendonck, Díaz, Rodríguez-Carvajal, Blanco, & Moreno-Jiménez, 2007; Fernandes, Vasconcelos-Raposo, & Teixeira, 2010; Kállay & Rus, 2014; Kitamura et al., 2004; Sirigatti et al., 2013, 2009; Villar, Triadó, & Celdrán, 2010). However, many of these applications vary widely on the number of items assessed. In many cases, it is not reported which or how items were selected for the short versions (e.g. Ryff & Keyes, 1995) and they often appear to differ across studies, making the comparison of measurement models troublesome at best. Nevertheless, the use of these shortened scales does indicate that there is a need for an adequately shortened version of the Ryff-Scale. In fact, three large US surveys - National Survey of Midlife in the US (MIDUS), National Survey of Families and Households (NSFH), and the Wisconsin Longitudinal Study (WLS) - employed a shortened version of the Ryff-Scale with three items per facet. While it is possible to determine which items were used in these surveys, information as to how these items were selected from the long version is not accessible.

3.3 Optimization Problem

In this section the properties of the optimization problem $(\mathcal{S}, f, \Omega)$ used in this evaluation are described in further detail. As discussed in Section 2.1, the optimization problem is given by the set of viable solutions $\mathcal{S}$, the quality function f, and the set of imposed constraints Ω. Because which solutions are contained in $\mathcal{S}$ is dependent upon Ω, the latter is discussed first.

As shown in Sections 1.3.2 and 2.1, representing item selection as an IMKAR results in three general constraints:

(ω_1) the sum of weights does not exceed capacity - Equation (1.10),

(ω_2) items are selected specifically in their respective facets - Equation (1.12), and

(ω_3) items may be selected to only one facet - Equation (1.13).

In a simplification of ω_1 for this specific application, all weights are set to be $w_{im} = 1$ and all facet capacities are set to $a_m = 3$, resulting in a short form consisting of 18 items.

Constraint ω_2 ensures that items can only be sampled to indicate the facet to which they pertain from a theoretical standpoint. Additionally, ω_3 ensures that each item is sampled to indicate only one latent variable in the short-form. As pointed out in Section 3.2, the items of the Ryff-Scale used in this evaluation were constructed in separate pools for each of the six facets, making an even stricter version a viable assumption about the item populations: it is assumed that all $\mathcal{C}_m$ are disjoint sets ($\mathcal{C}_m \cap \mathcal{C}_{m'} \equiv \emptyset$, for all $m \neq m'$), meaning that all items in the 54 item original are assumed each measure only a single facet. The item allocations are shown in Table 3.1.

Given the set of constraints Ω, the set of viable solutions $\mathcal{S}$ can be defined. This set consists of all possible short forms of the Ryff-Scale with 3 items per facet, items allocated only in line with the allocation key provided by Table 3.1, and items in the same order as they were in the original 54-item scale. Using Equation (2.5) this amounts to a total of

$$S = \prod_{m=1}^{M} \binom{I_m}{a_m} = 334569553920$$

viable solutions.

As pointed out in Section 2.1, the objective function f should be chosen in reference to the specific optimization problem at hand. In this evaluation f is given by

$$f(s) = \begin{cases} \Phi(s), & \text{if } s \in \mathcal{S}^* \\ 0, & \text{else} \end{cases}, \tag{3.3}$$

where $\Phi(s)$ is the pheromone function given in Equation (3.4) and $\mathcal{S}^* \subseteq \mathcal{S}$, for which the CFA returns a proper solution, meaning that the model converges and all model matrices are positive-definite.

For solutions $s \in \mathcal{S}^*$, the pheromone function is defined as:

$$\Phi(s) = \frac{1}{1 + e^{-10(crel_s - 0.4)}} + \left(.5 - \frac{.5}{1 + e^{-100(\text{RMSEA}_s - .05)}}\right) + \left(.5 - \frac{.5}{1 + e^{-100(\text{SRMR}_s - .05)}}\right). \tag{3.4}$$

with *crel* representing a composite reliability measure across all facets, RMSEA representing the Root-Mean-Squared-Error-of-Approximation, and SRMR representing the Standardized-Root-Mean-Residual of the solution s. The latter two components are each limited to .5, while reliability is limited to 1 to ensure that reliability and model fit are equally important in determining the quality of the solution. Figure 3.2 shows the solution quality as a function of RMSEA and

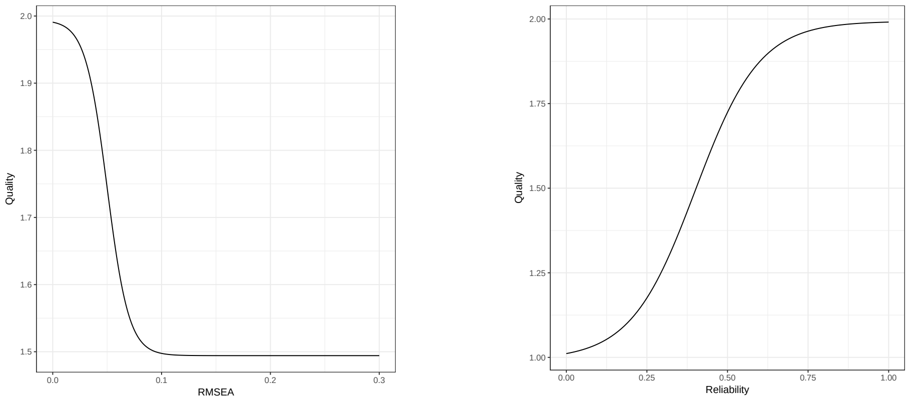


Figure 3.2: Solution quality as a function of RMSEA and composite Reliabililty.

composite reliability when the other two components included in Equation (3.4) are held at their respective optimum. Due to the additive nature of Φ the value of the other components is of no consequence for the relationship between a component and the quality, however. The function of the SRMR is not shown, because it is identical to that of the RMSEA. Overall the function was set in this manner to allow for maximum discrimination at the values of .05 for both SRMR and RMSEA and to reward all models with fit worse than .1 on both indexes with a value of approximately 0 for this component of the solution quality.

Via the definition of $\Phi(s)$ it is possible to derive $\min \Phi(s)$ and $\max \Phi(s)$ as the theoretical bounds on the pheromone function. Both of these can be determined in a straightforward manner, because all components are known. Reliability, as a variance component, has the range $[0; 1]$. The RMSEA has a lower bound of 0 and no theoretical upper limit, while the range of the SRMR is also $[0; 1]$ (West, Taylor, & Wu, 2012). With the minimum possible value for reliability ($crel = 0$) and the maximum possible value for model misfit (RMSEA $= \infty$, SRMR $= 1$), $\min \Phi(s) = 0.018$ can be determined as the theoretical lower bound of $\Phi(s)$. The upper bound, $\max \Phi(s) = 1.991$, is obtained when reliability is maximal ($crel = 1$) and both indicators of model misfit are minimal (RMSEA $= 0$, SRMR $= 0$). The lower bound of the quality function f is actually $\min f = 0$, due to the conditional statement in Equation (3.3).

3.4 Referential Solutions

Given the optimization problem just described, it is possible to construct some solutions for reference and evaluate them using the quality function f. This is done to understand the performance of each replication of the evaluation more fully, by setting reference standards against which to compare solutions.

The obvious first choice is the analysis of the short form used in the large-scale surveys.

Table 3.2: Results of the CFA with the short-form of the Ryff-Scale used in survey studies.

					Latent Correlations					
Subscale	Items			Rel.	1	2	3	4	5	6
Self-Acceptance	4	23	31	0.660	1.000					
Positive Relations	5	34	39	0.458	0.729	1.000				
Autonomy	25	35	52	0.667	0.608	0.392	1.000			
Environmental Mastery	2	7	17	0.473	0.822	0.610	0.520	1.000		
Purpose in Life	8	42	46	0.532	0.409	0.276	0.289	0.418	1.000	
Personal Growth	21	45	50	0.538	0.762	0.534	0.553	0.637	0.362	1.000

Table 3.3: Results of the CFA with the short form of the Ryff-Scale determined in the optimal solution s^{opt}.

					Latent Correlations					
Subscale	Items			Rel.	1	2	3	4	5	6
Self-Acceptance	9	23	28	0.683	1.000					
Positive Relations	10	24	39	0.646	0.458	1.000				
Autonomy	6	19	52	0.645	0.736	0.310	1.000			
Environmental Mastery	20	36	49	0.568	0.560	0.476	0.487	1.000		
Purpose in Life	13	27	30	0.507	0.476	0.575	0.310	0.625	1.000	
Personal Growth	3	50	54	0.453	0.621	0.476	0.492	0.616	0.799	1.000

Because this short form was not constructed with the goal of optimizing the measurement CFA in mind, it should not be expected to perform optimally with regards to the quality function used in this section. Applying a six-factor CFA with freely correlated facets to the selection of items used in the surveys produces a Heywood-Case in this dataset. This was due to a negative residual variance pertaining to item 42. The model shows moderate fit to the data ($\chi^2 = 808.006$, $df =$ 120, $p < .001$, RMSEA = 0.062, SRMR = 0.047, CFI = 0.855). Table 3.2 provides a detailed overview of the results of the CFA regarding this particular short form. The reliability estimates in the fifth column were obtained using the SEM reliability computation reported by Yang and Green (2010). The results indicate somewhat mediocre reliabilities of the facets, ranging from 0.458 to 0.667, but an adequate reliability of the entire scale at 0.892. The latent correlations do not seem to indicate support for the claim of non-discernible factors, though it should be noted that in contrast to the results reported by Springer and Hauser (2006), this model does not include a method factor to control for the phrasing effect of the items.

Following the definition of the quality function in Equation (3.3), the determined quality of this solution is $f(s^{ts}) = 0$, because the negative residual variance of item 42 indicates that $s^{ts} \notin \mathcal{S}^*$. The superscript ts is used to indicate that this is the solution derived via *t*heory-guided *s*election. Disregarding this condition, the pheromone function returns $\Phi(s^{ts}) = 1.396$ as a possible reference value for the overall quality of this solution.

Table 3.4: Results of the CFA with the short form of the Ryff-Scale determined in the median solution s^{md}.

					Latent Correlations					
Subscale	Items			Rel.	1	2	3	4	5	6
Self-Acceptance	9	31	48	0.691	1.000					
Positive Relations	5	15	39	0.481	0.713	1.000				
Autonomy	6	19	35	0.638	0.633	0.366	1.000			
Environmental Mastery	2	12	29	0.433	0.887	0.803	0.447	1.000		
Purpose in Life	13	33	38	0.605	0.831	0.583	0.496	0.894	1.000	
Personal Growth	41	50	54	0.393	0.729	0.672	0.514	0.682	0.766	1.000

Perhaps the most important solution for the evaluation of the `stuart` approach is the optimal solution s^{opt}. This solution is of such central importance because it can be seen as the target solution each run of the algorithm sets out to find. Therefore, comparing the results of any single replication against s^{opt} can be used as an indicator of how close to the goal finding the optimal solution any single instance is. Because this solution is not known, it will be assumed that the best solution found throughout the entirety of the evaluation study (which generated just shy of 342 million solutions) is the optimal solution. Table 3.3 provides an overview of the results of this optimal solution. The reliabilities found for the subscales are similar in magnitude to those determined for s^{ts}. The composite reliability of the entire scale was 0.924 for this solution. As should be expected, the model fit the data quite well ($\chi^2 = 188.168$, $df = 120$, $p < .001$, RMSEA $= 0.019$, SRMR $= 0.021$, CFI $= 0.986$). For this solution the quality can be determined to be $f(s^{opt}) = 1.947$.

Another approach of obtaining referential solutions is to construct a random sample of viable solutions and determine the quality distribution. The basic idea underlying this approach is that drawing a random sample from the set of possible solutions constitutes the null algorithm. This algorithm is defined solely by random chance and can therefore be interpreted as the null hypothesis against which any stochastic algorithm should be compared (c.f. Eiben & Smith, 2015). Generating 10000 random solutions results in 5802 solutions with $f(s) = 0$, due to non-positive definite latent variance-covariance matrices or negative residual variances. The distribution of $\Phi(s)$ for the remaining 4198 is shown in Figure 3.3, with the dashed line indicating $\Phi(s^{ts})$ and the solid line representing $\Phi(s^{opt})$. For these solutions, the average quality is $M_f = 1.481$ with $SD_f = 0.122$. Assuming a normal distribution for $\Phi(s)$, the chances of finding solutions as good or better than the two referential solutions discussed here are $p[\Phi(s) \geq \Phi(s^{ts})] = 0.759$ and $p[\Phi(s) \geq \Phi(s^{opt})] < .001$. With skew$_\Phi = 0.36$ and kurt$_\Phi = 1.175$, assuming a normal distribution seems to be a somewhat serviceable approximation for this rough placement of these solutions in the spectrum of possible solutions.

The third and final referential solution that will be of relevance in this section is the median

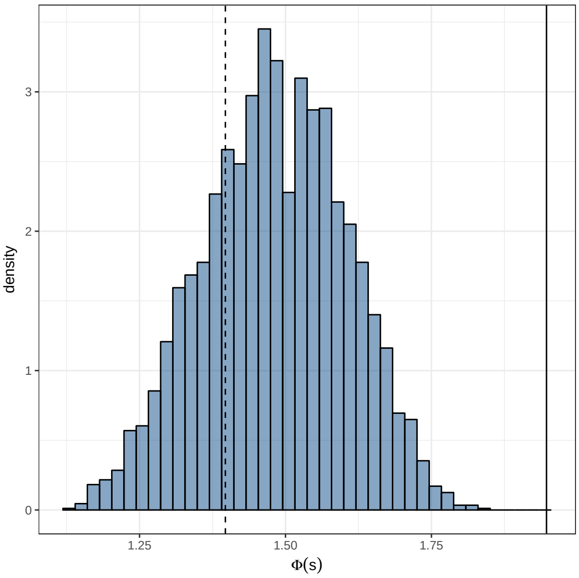


Figure 3.3: Histogram of the quality of the random solutions. The dashed line indicates the pheromone of the theory-guided selection, the solid line the pheromone of the optimal solution.

solution s^{md} of the random sample. This solution represents the short form that leads to the median pheromone among all proper solutions - i.e. solutions that do not result in $f(s) = 0$ due to Equation (3.3). The composite reliability of the entire scale of this solution is 0.908. The model does not fit the data adequately ($\chi^2 = 731.622$, *df* $= 120$, $p < .001$, RMSEA $= 0.058$, SRMR $= 0.043$, CFI $= 0.887$). For this solution the quality can be determined to be $f(s^{md}) = 1.483$. The detailed results of this solution are presented in Table 3.4.

While this particular solution is of no real value from a substantive point of view, it will be used to provide heuristic information indicating a sub-optimal solution in the evaluation. This solution is used instead of the theory-guided solution because (a) s^{ts} leads to an improper solution in the dataset used here, and (b) an item allocated to the facet personal growth in the short-form used in survey studies is actually part of the facet self-acceptance in the German version used in this study (item 45).

The consequence of using heuristic information leading to improper solutions is unknown and not investigated in detail in this thesis. However, a consequence of using such a solution would be that the solution favored in the heuristic information is unable to deposit pheromone because $f(s) = 0$, in such cases. Therefore, only other solutions would be able to deposit pheromone on the decision nodes, leading to a tug-o-war between the two components $\phi_{ij}(t)$ and η_{ij} in Equation (1.16), when determining the selection probability of items, i.e. this would require the pheromones to “overcome” the selection bias provided by the heuristics. It seems more informative to investigate the behavior of the algorithm when the construction of solutions in line with the heuristic information can lead to pheromone deposit, because in this case different parameter settings can lead to the heuristic information becoming dominant over the information of Φ or the heuristic information becoming irrelevant. As shown in Tables 3.3 and 3.4 the optimal and median short-forms share 7 Items, with only the subscale pertaining to environmental mastery showing no overlap between the two solutions.

3.5 Perfomance Measures

Throughout this study, the evaluation of the performance of the `stuart` approach will most often be determined by the comparison against the optimal solution s^{opt}. This is done because s^{opt} constitutes the best possible solution to the problem defined in Section 3.3. In line with the recommendations by Birattari and Dorigo (2007) for evaluating algorithms, measures of the average performance will be most prevalent in this section and only in a few select cases will the best or worst performance across replications be of interest.

The performance measures used in this evaluation are in line with those proposed by Eiben and Smith (2015) for the evaluation of evolutionary algorithms. The first two measures discussed are measures of the effectiveness of the algorithm, namely measures that indicate how well the final solution fulfills the set criteria.

The first performance measure of interest is the success rate (SR), defined simply as the relative frequency with which the optimal solution can be found for any given parameter constellation. The value of this measure is obvious: higher SRs of parameter constellations imply a better chance of finding the truly optimal solution in any single application. As stated above, the optimal solution is not known and it is assumed that the best solution that was found across all conditions of the evaluation is the optimal solution, making it possible to determine SR.

However, given the vast amount of possible solutions, it is often either not possible or not necessary to find the optimal solution. In these cases, it is of relevance to know whether the final solution is good enough - i.e. if it fulfills the set criteria well enough by providing a solution with a very high value of $f(s^{gb})$. For this Eiben and Smith (2015) propose using the mean best fitness (MBF), defined simply as the mean of $f(s^{gb})$ over all replications. The value used in this evaluation differs slightly and will be called relative deviation (RD). It is obtained via

$$RD = \frac{\frac{1}{R}\sum_{r=1}^{R} f(s_r^{gb}) - f(s^{opt})}{f(s^{opt})}, \tag{3.5}$$

where R is the number of replications. The RD is used instead of the MBF because it is more readily interpretable without deeper knowledge of the quality function and its behavior. As a measure of relative deviation from the optimal solution it is scaled from -1 to 0, with -1 occurring only when $f(s^{gb}) = 0$ (i.e. when no viable solution is found) and 0 being reached if the optimal solution is recovered.

The RD can also be computed with the regard to other referential solutions, such as s^{md}, though this is of little consequence when comparing the performance of conditions, because different variants of RD are simply linear transformations of each other. The relative deviation regarding the median solution, for example, can be computed from the RD regarding the optimal solution as

$$RD^{md} = \frac{f(s^{md}) - f(s^{opt})}{f(s^{opt})} + \left(\frac{f(s^{md}) - f(s^{opt})}{f(s^{opt})} + 1 \right) RD. \tag{3.6}$$

In most cases the RD will be computed with regard to the optimal solution. All others cases will be made explicit.

It should be noted that RD and SR are likely to be correlated, but provide different perspectives on the performance of the algorithm. While the SR is a rather strict assessment of performance, because it views any replications not ending with the optimal solution as failures, the RD is a more fine grained measure that can provide a lot of additional information. Especially in situations in which there is a wide array of very good solutions, a low SR is not necessarily indicative of poor algorithmic performance. The opposite may also be the case - though unlikely in this situation. It is possible for an algorithm to detect the optimal solution often, but fail to provide a usable solution in cases in which the optimum is not found.

In addition to these two measures, which relate more or less directly to the optimal solution, selection consistency SC can be defined as the proportion of the items favored by the heuristics, which are present in the final solution.[2] This provides an indicator of the integration of the heuristic information into the construction of subtests.

A fourth measure used in the evaluation of the performance of the `stuart` approach under different circumstances, does not investigate the quality of the final solution, but is concerned with the time it takes for a replication to converge or reach the abort criterion. Because the majority of runtime is spent in the evaluation of the solutions (i.e. in the estimation of the CFAs) and the actual time spent is determined by a multitude of factors outside of the algorithm's reach (e.g. raw processing power, CFA estimation software optimizations) the total number of ants that are run (absolute runtime; AR) will be used as an indicator of runtime.

In addition to these global performance measures, two measures can be defined as functions of time. The first is the time-sensitive version of the relative deviation defined in Equation (3.5), which is simply the $RD(t)$: the RD evaluated after any arbitrary number of colonies t. In this sense, the RD is a special case of the $RD(t)$ with $t = T$.

The second measure is the relative exploration $RE(t)$ of a colony t. This can be operationalized as the proportion of solutions constructed in colony t that are proper and unique solutions. RE can then be defined as the mean $RE(t)$ over all t. Thus, the RE represents the overall

[2] Due to the design of this evaluation the heuristic information favors exactly 18 items - as many as are selected in the final solution. While this is not necessary in standard usage of the approach, it has the benefit of scaling SC from 0 to 1.

proportion of unique solutions constructed over the course of the entire run.

The purpose of both of these time-dependent measures is to guide decisions regarding the settings of scheduled parameters in the second part of this evaluation. As pointed out above, it is the aim of these parameter schedules to capitalize on the effects of certain parameter settings, to ensure a wide exploration of the space of possible solutions during the early stages of the run and a quick exploitation of good solutions in the later stages of the run.

Dorigo and Gambardella (1997) proposed a more finely grained performance measure for ACO-based search algorithms called the λ-branching factor. This measure gives extremely detailed information about the search behavior of an algorithm by providing a summary statistic of the search space based on each single decision node. Dorigo and Stützle (2004) propose using the average entropy as more readily interpretable alternative to λ. However, both of these statistics are very expensive to compute because they require an additional computational step and data storage for each decision node after each iteration. Because of this, both of these approaches are not used in this thesis.

To express the effects experimental factors have on these performance measures in a standardized and easily interpretable fashion, Cohen's d will be used to contrast two specific parameter constellations - most often when investigating the effects of the deposit rule on the performance measures. These estimates and their 95% confidence intervals will be computed using the `effsize`-Package (Torchiano, 2016). For most instances, however, ΔR^2 will be used. This specific ΔR^2 is determined as the difference in R^2 of the OLS regressions with and without the relevant factor and all its interactions with other independent variables. In these settings all independent factors are treated as nominal-scaled and are dummy-coded for inclusion in the regression. The 95% confidence intervals for these ΔR^2 are computed in accordance to Cohen, Cohen, West, and Aiken (2003, p. 88).

3.6 Constant Parameter Settings

As discussed above, the evaluation is performed in a two-step procedure with the first step being a fully crossed design of several constant parameter settings. This section will report the settings of the parameters in this specific evaluation - for a more detailed discussion of the expected effect of these parameters on the overall performance of the search algorithm, please see Section 3.1.

The first parameter described in Section 3.1 is the heuristic information provided in $\mathbf{H}$. In this specific evaluation the influence of $\mathbf{H}$ is investigated in three different settings. The first is the complete absence of heuristic information. In this case all elements of $\mathbf{H}$ are set to 1, making the selection of items dependent only on the pheromones deposited. The second setting is the optimal setting, in which the heuristic information favors the items selected in the optimal solution s^{opt}. The third setting favors the median solution s^{md}, as described in Section 3.4. In both of the latter cases the items of the "favorable" solution are weighted with 2 in $\mathbf{H}$, while all others are weighted with 1, making the selection of of a favored item 2^{β} as likely as that of an item not contained in s^{opt} or s^{med}, respectively, at the beginning of the search. These three settings for the heuristics are used for several reasons. Choosing no heuristics is most likely a very common case in practice, thus making an evaluation of the performance of the approach necessary under these conditions. When heuristics are provided, they are most likely indicative of a solution close to optimal. Therefore, they should bias the search towards an area with very good solutions. Optimal and median quality heuristics are used as extreme examples here to evaluate the performance of the approach in both of these cases. In applications, the quality of heuristics should indicate solutions somewhere between these two extremes.

The non-linearity coefficient of the pheromone α is varied in four steps: 1, 1.5, 2.5, and 5. As described above, it is expected that lower values of α lead to more exploration but less exploitation - meaning a slower convergence to a better solution. For the non-linearity coefficient of the heuristic information β, the same four steps were chosen. For those conditions in which heuristic information is included, the settings of α and β were fully crossed, while for the condition including no heuristic information $\beta = 1$ is used.

The evaporation coefficient ρ is varied in three steps: .95, .8, and .5. Higher values are expected to lead to a slower convergence to a better solution. Due to the fact that the range between ϕ^{min} and ϕ^{max} is dependent upon the values of ρ - as shown in Equations (1.22) and (1.23) - the tolerance parameter included in the convergence criterion stated in Equation (2.11) is set to be

$$tol = \frac{\max f(s)}{50(1-\rho)}, \tag{3.7}$$

which simply states that the tolerance is $\frac{1}{50}$ of the maximum of pheromone that can be deposited on any given element of the pheromone matrix. This is similar to ϕ^{max} as defined in equation (1.22) but does not pertain to any actual optimal or global best solution, but instead to the theoretical upper bound of $f(s)$ - calculated in Section 3.3 via $\max f(s) = \max \Phi(s) = 1.991$.

The number of iterations K is kept constant for all conditions. Throughout the entire evaluation this abort criterion is set to 256 colonies after the last global best solution s^{gb} was found.

The number of ants per colony is varied in three steps: 32, 16, and 8. As noted in Section 3.1.6, higher values of K lead to less cooperation between ants, thereby leading to a wider search and slower convergence.

The value of $p_{s^{gb}} = .005$ is the same across all conditions, because the size of the problem does not change. In this case $p_{s^{gb}}$ is roughly 1.67×10^9 times the probability of finding the global best solution by chance.

Finally, both deposit rules - iteration best and global best pheromone deposit - are used in this evaluation. In line with Stützle (1998) it should be assumed that the *ib* deposit rule performs better, because the `stuart` approach does not include local search.

Table 3.5: Overview of the constant parameter settings used in the fully-crossed evaluation design for Evaluation 1. Deposit rules are shortened *ib* for iteration-best and *gb* for global-best.

α	1	1.5	2.5	5
β	1	1.5	2.5	5
ρ	.5	.8	.95	
K	8	16	32	
Deposit Rule	*ib*	*gb*		
H	none	optimal	sub-optimal	

These different parameter settings result in a total of 648 different conditions. Table 3.5 provides an overview of the evaluation conditions. Each of the 648 conditions was replicated 100 times.

With these restrictions in place, it is possible to derive the minimum and maximum runtime - i.e. the theoretical bounds of AR. Theoretically, AR has no upper bound unless a set number of colonies is defined as an abort criterion. Given the conditions in this evaluation the maximum runtime is given by

$$\max AR = 256SK \tag{3.8}$$

with S being the number of possible solutions computed in Equation (2.5) as 334569553920, K being the number of ants in a given condition and 256 being the number of colonies defined as the abort criterion. This upper limit of AR is reached only if all possible combinations have unique fitness and are found as the s^{gb} in sequential order of their $f(s)$, always in the 256th colony after finding the previous solution.

Using the minimum required iterations presented in Equation (3.2), the lower bound on AR - and thus the minimum number of ants required before convergence - can be computed as:

$$\min AR = \left\lceil \frac{\log\left(\frac{\phi^{min}+tol}{\phi^{max}}\right)}{\log(\rho)} \right\rceil K. \tag{3.9}$$

With these parameter settings there are nine different lower bounds for AR in this study. When $\rho = .5$, the minimum number of iterations required to reach the convergence criterion is $t = 5$, thus $\min AR = 5K$. For conditions with $\rho = .8$ this is $\min AR = 13K$ and with $\rho = .95$ it is $\min AR = 52K$.

Additionally, the probability of constructing the solution favored by the heuristic information in the initial step can be computed. During the first colony ($t = 1$) all items have the the same pheromone (because all are initialized to the same value) and thus the selection probability given by Equation (2.6) simplifies to

$$p(x_{im} = 1|t) = \frac{[\eta_{im}]^\beta}{\sum\limits_{i=1}^{I_m} [\eta_{im}]^\beta}. \tag{3.10}$$

Therefore, the probability of selecting only those items stemming from the heuristically favored solution s^* can be computed as

$$p(s = s^*) = \prod_{m=1}^{M} \prod_{i=1}^{a_m} \frac{i2^\beta}{i2^\beta + (I_m - a_m)} \tag{3.11}$$

during the first iteration, when using $\eta_i = 2$ for $i \in s^*$ and $\eta_i = 1$ for $i \notin s^*$. With this, the probability of generating the favored solution at least once in the initial colony can simply be computed as $1 - \{1 - (p[s = s^*])\}^K$.

This is of relevance, because for some settings it becomes extremely likely to generate the solution that is favored by the heuristics. In the most extreme case of this evaluation (for $\beta = 5$ and $K = 32$), this probability is 0.993. Due to the nature of the `stuart` approach, if the heuristically favored solution is constructed during this initial iteration, the final solution cannot be worse than the heuristically favored solution. While this is a good property of the approach in general, it means that there is likely a considerable proportion of replications in the conditions with optimal heuristics that encounter the optimal solution in the first iteration and will then run unnecessarily long.

3.7 Results for the Constant Parameter Settings

The results for this part of the parameter evaluation will be presented in three sections. The first is dedicated to the results derived from the 72 conditions for which no heuristic information was provided. The following sections will then focus on those conditions in which optimal and sub-optimal heuristics were provided.

3.7.1 No Heuristic Information

Errors In total 28 of 7200 replications (0.389%) did not terminate with either of the two predefined abort criteria. All of these errors were due to none of the solutions in the first colony of the run providing a proper solution - i.e. all ants of the first colony provided $f(s) = 0$. As can be expected, this was exclusively the case in conditions with $K = 8$.

Success Rate (SR) Over all conditions the success rate was $SR = 0.039$ indicating a rather low recovery rate of s^{gb} over a wide array of parameter settings. Table 3.6 provides a bit more detail about the condition specific performance regarding the SR. None of the replications using $\alpha = 5$ resulted in the optimal solution, and all conditions with $\alpha = 2.5$ performed to $SR \leq .05$,

Table 3.6: Condition-specific SR for cases with no heuristic information.

		ib-Deposit				*gb*-Deposit			
K	ρ	$\alpha = 1$	$\alpha = 1.5$	$\alpha = 2.5$	$\alpha = 5$	$\alpha = 1$	$\alpha = 1.5$	$\alpha = 2.5$	$\alpha = 5$
	0.5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
8	0.8	0.000	0.010	0.000	0.000	0.000	0.000	0.000	0.000
	0.95	0.020	0.020	0.000	0.000	0.052	0.091	0.010	0.000
	0.5	0.030	0.000	0.000	0.000	0.000	0.000	0.000	0.000
16	0.8	0.160	0.030	0.000	0.000	0.020	0.020	0.000	0.000
	0.95	0.140	0.140	0.000	0.000	0.100	0.130	0.020	0.000
	0.5	0.030	0.060	0.000	0.000	0.000	0.050	0.000	0.000
32	0.8	0.480	0.120	0.000	0.000	0.040	0.110	0.010	0.000
	0.95	0.160	0.200	0.050	0.000	0.280	0.190	0.040	0.000

indicating lower success rates for larger degrees of non-linearity of the pheromone in determining selection probability. In cases with $\alpha = 1$ or $\alpha = 1.5$, the SR is almost always higher for the iteration-best deposit rule than for the global-best deposit rule and generally seems to increase with K. With $\alpha = 1.5$, larger values of ρ lead to higher SR, while the effect of ρ does not seem to follow a simple tendency when $\alpha = 1$. For these conditions, $\rho = .8$ resulted in the highest SR (peaking at $SR = 0.48$). When using the global-best deposit rule, however, conditions with $\rho = .95$ outperformed the other two parameter settings with regards to sucessfully recovering s^{opt}.

Relative Deviation (RD) The average RD across all parameter settings is -0.037, indicating that the average final solution performed approximately 4% worse than the optimal solution with regards to the quality function. In contrast, the average RD when using the s^{md} described in Section 3.5 as the reference solution is 0.265, indicating that the average best solution outperforms the median solution by approximately 26%. As was the case with the SR the overall performance regarding RD is much less informative than the comparisons between the conditions. Figure 3.4 shows the Boxplots of the RD for each of the conditions incorporating no heuristic information.

Generally, higher settings of α lead to worse RD values - as indicated by the downward shift of the Boxplots across the panels of Figure 3.4 from left to right. In total, the variations in α accounted for a $\Delta R^2 = 0.465$ (95% $CI = [0.445; 0.484]$) and represent the most important factor in determining the values of RD.

Similarly, larger values of K consistently lead to better final solutions ($\Delta R^2 = 0.095$, 95% $CI = [0.081; 0.109]$). The deposit-rule does not make a tangible difference for RD in cases in which α is either 2.5 or 5 (Cohen's $d_{\alpha=2.5} = 0.026$, 95% $CI = [-0.066; 0.119]$, Cohen's $d_{\alpha=5} = -0.011$, 95% $CI = [-0.103; 0.082]$). In cases with $\alpha = 1.5$ these differences are also negligible (Cohen's $d_{\alpha=1.5} = 0.128$, 95% $CI = [0.035; 0.221]$). However, when $\alpha = 1$ the iteration-best deposit rule outperforms the global-best deposit rule quite considerably ($d = 0.813$, 95% $CI = [0.717; 0.909]$).

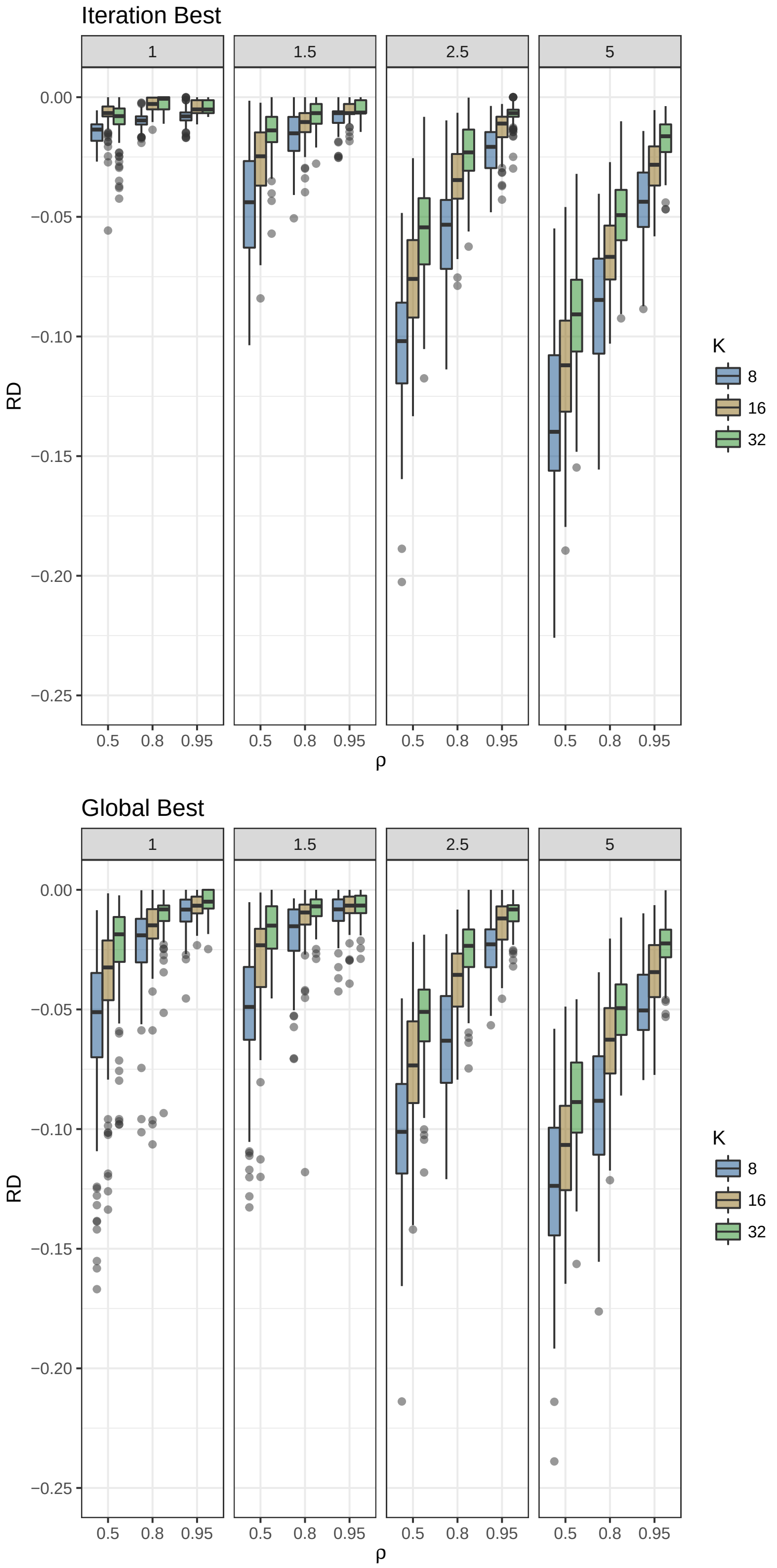


Figure 3.4: Boxplots of the relative deviation (RD) of the final solutions from the optimal solution. The columns of panels depict different values of α.

Table 3.7: Average and worst-case *RD* for the four best performing conditions. The ranks pertain to the total of all 72 conditions without heuristic information.

				Average		Worst-Case			
Deposit	α	K	ρ	RD	Rank	RD	Rank	SR	AR
ib	1	16	0.8	-0.003	2	-0.014	4	0.160	9082.560
ib	1	16	0.95	-0.004	4	-0.011	3	0.140	10790.400
ib	1	32	0.8	-0.002	1	-0.011	2	0.480	13342.080
ib	1	32	0.95	-0.004	3	-0.008	1	0.160	16690.560

As was the case for the SR, there is a general tendency for ρ to increase values of RD ($\Delta R^2 = 0.343$, 95% $CI = [0.324; 0.362]$), except when using the iteration-best deposit rule with $\alpha = 1$, where $\rho = .8$ (in combination with $K = 32$) leads to the best average performance of all conditions ($M_{RD} = -0.002$). However, while this constellation leads to the best *average* performance, its *worst-case* performance is worse than that of the condition with $\alpha = 1$, $K = 32$, and $\rho = .95$. Table 3.7 shows the four best performing conditions with regards to average and worst-case performance and the respective ranks amongst all 72 conditions investigated in this section. Only the best four are presented because they are the same in both situations.

For reference, the worst final solution of any of these four conditions (i.e. the worst final solution found when using the iteration-best deposit rule, $\alpha = 1$, $K = 16$, and $\rho = .8$) resulted in $f(s) = 1.921$. This quality value was derived from a shortened scale with a composite reliability of 0.896 and a measurement model with good overall fit (RMSEA = 0.025, SRMR = 0.024, CFI = 0.968). Under the normal distribution assumed for the 10000 random samples described in Section 3.4 this solution (s^{gb}) provides $p[\Phi(s) \geq \Phi(s^{gb})] < .001$.

Absolute Runtime (*AR*) On average, 2317.405 CFAs were run per condition. Again, this number is not as informative as the condition specific runtimes, for which the Boxplots are shown in Figure 3.5. The boxplots indicate general increases in runtime for larger values of K and ρ, as well as smaller values of α.

As was the case for RD, the deposit rule does not have a substantial impact on AR in conditions with non-linear deposit rules (Cohen's $d_{\alpha=1.5} = 0.128$, 95% $CI = [0.035; 0.221]$, Cohen's $d_{\alpha=2.5} = -0.053$, 95% $CI = [-0.146; 0.04]$, Cohen's $d_{\alpha=5} = 0.181$, 95% $CI = [0.088; 0.273]$). However, conditions in which the pheromone is deposited in a linear fashion, the global-best deposit rule reaches either of the abort criteria much faster ($d = 1.426$, 95% $CI = [1.322; 1.53]$).

Overall, the most important determinant in the absolute runtime was the non-linearity coefficient α, with larger values leading to much shorter runtimes ($\Delta R^2 = 0.496$, 95% $CI = [0.478; 0.515]$). For both, ρ ($\Delta R^2 = 0.324$, 95% $CI = [0.307; 0.341]$) and K ($\Delta R^2 = 0.196$, 95% $CI = [0.183; 0.21]$) larger values lead to considerably longer runtimes.

The fastest convergence observed was after just 40 ants - a feat that was accomplished 52

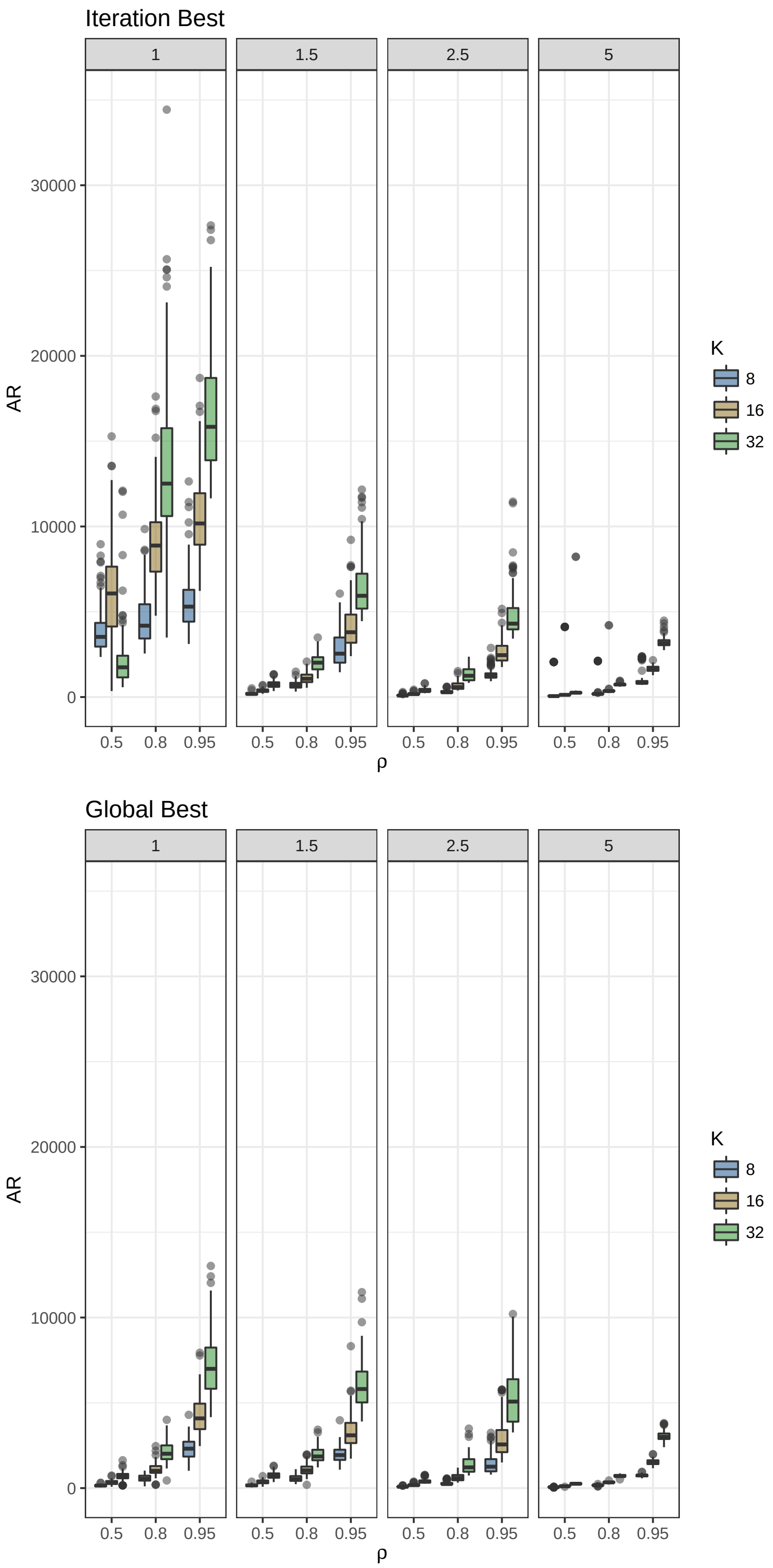


Figure 3.5: Boxplots of the condition specific absolute runtimes (AR). The columns of panels depict different values of α.

times. All of these replications used $\rho = .5$ and $K = 8$, for which 40 also constitutes the theoretical minimum required for reaching the convergence criterion as computed via Equation (3.9). Among these fastest convergences were conditions utilizing any setting for α in combination with the global-best deposit rule, but only conditions with $\alpha = 5$ when using the iteration-best deposit rule. Overall, 82 replications achieved minimum AR, with 79 (96.341%) of these stemming from conditions with $\rho = .5$, and the remaining replications stemming from conditions with $\rho = .8$. Additionally, 11 (13.415%) of these utilized the iteration-best deposit rule. The best of the 82 replications with minimal runtime ended with $RD =$ -0.046, while the worst ended with $RD =$ -0.214.

The slowest condition ended at the abort criterion after 34432 ants. As visible in Figure 3.5, this replication stems from the condition with the iteration-best deposit rule, $\alpha = 1$, $K = 32$, and $\rho = 0.8$. This replication ended in the optimal solution. It is noteworthy that, even though 34432 is quite a large number of CFAs to run, it represents running just 0.00001% of the total number of viable solutions.

Relative Exploration (RE) Figure 3.6 shows the condition specific Boxplots for RE. Over all conditions the average RE was 0.449, indicating that well under 50% of all constructed solutions were proper, unique solutions. Higher values of the non-linearity coefficient α lead to less exploration, as did higher values of K.

The findings concerning the evaporation coefficient ρ are mixed. When using the iteration-best deposit rule ρ had no overall effect on the degree of exploration ($\Delta R^2 = 0.003$, 95% $CI = [-0.001; 0.008]$). However, with the global-best deposit rule, higher values of ρ lead to considerably less exploration ($\Delta R^2 = 0.014$, 95% $CI = [0.006; 0.022]$). As visible in Figure 3.6, conditions with $\alpha = 1$ did not exhibit a clear trend.

Figure 3.6 shows that α and its interactions are extremely important in determining the degree of exploration across all conditions ($\Delta R^2 = 0.922$, 95% $CI = [0.916; 0.928]$). For conditions with $\alpha = 1$ and the iteration-best deposit rule the average RE is 0.827, and for conditions with $\alpha = 1$ and the global-best deposit rule the average is slightly lower, at $M_{RE} = 0.735$.

Relative Deviation over Time ($RD[t]$) Figure 3.7 shows the optimization history with regards to $RD(t)$ of each of the replications using the iteration-best deposit rule. As described above, absolute runtimes differed drastically between conditions with varying values for α, which is why the four panels of Figure 3.7 use different scales on the x-axis. For $\alpha = 1$ the quality of the iteration-best solution slowly approaches RD-values close to zero, finding the final solution after an average of 271.153 iterations and then replicating that final solution 218.036 times, on average. It should be noted, that not a single replication reached the abort criterion defined via the pheromone limits when using the iteration-best deposit rule, $\alpha = 1$, and $\rho = .95$. Instead, all of these replications reached the maximum number of iterations after the last improvement

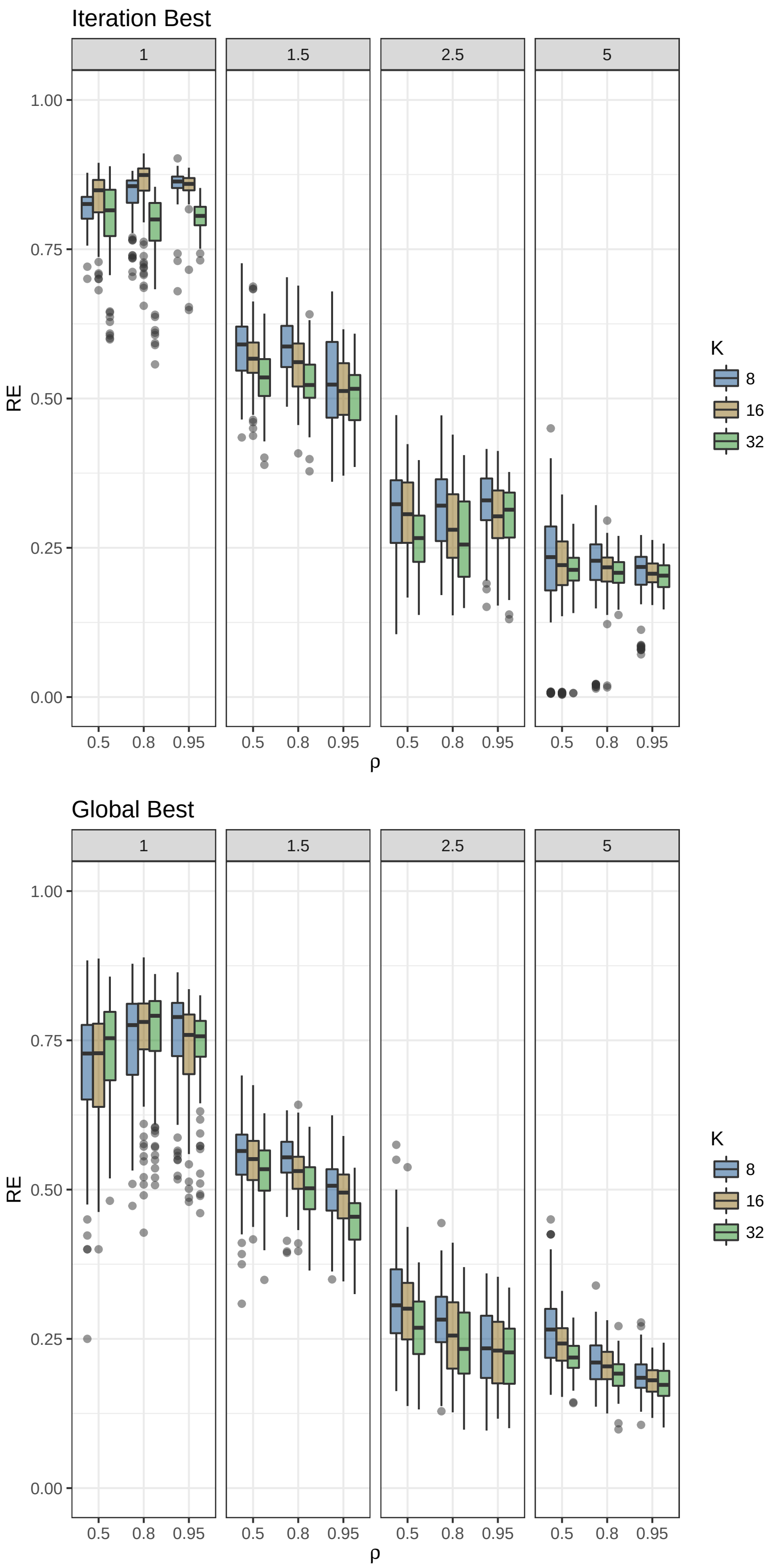


Figure 3.6: Boxplots of the condition specific relative exploration (RE). The columns of panels depict different values of α.

on s^{gb}.

Figure 3.7 shows that the decreased runtimes associated with larger values of α are due to quick convergence once an acceptable solution is found. In conditions with $\alpha = 5$ and $\rho = .5$ this quite often lead to convergence in a solution found in the first iteration, resulting in values of AR at the lower theoretical bound.

The drastic differences in behavior for different levels of α are not quite as pronounced in Figure 3.8, which depicts the $RD(t)$ of all replications using the global-best deposit rule. This is partially due to faster convergence with $\alpha = 1$, when compared to conditions using the iteration-best deposit rule. For these conditions, an average of 85.631 iterations was needed before the final solution was found, which is markedly faster than in the conditions utilizing the iteration-best deposit rule and $\alpha = 1$ ($d = 1.219$, 95% $CI = [1.118; 1.32]$). Additionally, the final solutions were replicated much less often ($M_{rep} = 30.877$, $d = 2.942$, 95% $CI = [2.808; 3.076]$), leading to the results concerning the differences in AR described earlier.

Relative Exploration over Time ($RE[t]$) Figure 3.9 shows the LOWESS lines for the relative exploration over time. With the exception of conditions with $\alpha = 5$ and $\rho = .8$ or $\rho = .5$, all conditions experienced an initial burn-in phase during which the $RE(t)$ values increased. This is mostly due to a large amount of improper solutions during the early exploration phase.

Defining the burn-in phase as the phase until the 90th percentile in $RE(t)$ is reached, Table 3.8 shows the average number of iterations needed for the burn-in phase. Across all conditions, lower values of ρ lead to a faster burn-in, as did lower values for K. Higher values of α had the same effect, but the $RE(t)$ they achieved at the 90th percentile was much lower. Overall, conditions utilizing $\alpha = 1$ with the iteration-best deposit rule all showed 90th percentiles of $RE(t) > .90$, while those using $\alpha = 1$ and the global-best deposit rule achieved only marginally worse 90th percentiles with all $RE(t) > .88$. As is visible in Figure 3.9, conditions with higher values of α achieved much worse performance with regard to the 90th percentile, with the general trend of lower $RE(t)$ accompanying lower values of ρ. This is not the case in conditions with $\alpha = 1$, where lower values of ρ actually lead to higher 90th percentile $RE(t)$.

As already mentioned above, the average RE is lower for higher values of α. The most notable pattern is that - independent of the deposit rule used - values of $\alpha = 1$ lead to consistently high $RE(t)$ values after the initial burn-in phase, with only the two conditions ($K = 32$ and $\rho = .8$ or $\rho = .95$) showing signs of decline in $RE(t)$ during later phases. For conditions with $\alpha = 1.5$ or $\alpha = 2.5$, the burn-in is followed by a short peak in $RE(t)$ before a quick decline leads to convergence in solution.

Summary In situations with no heuristic information there is a general tendency for parameter constellations with smaller values of α, as well as larger values of ρ and K to lead to slower but markedly better solutions. Overall, $\alpha = 1$ showed the most promising results regarding solution

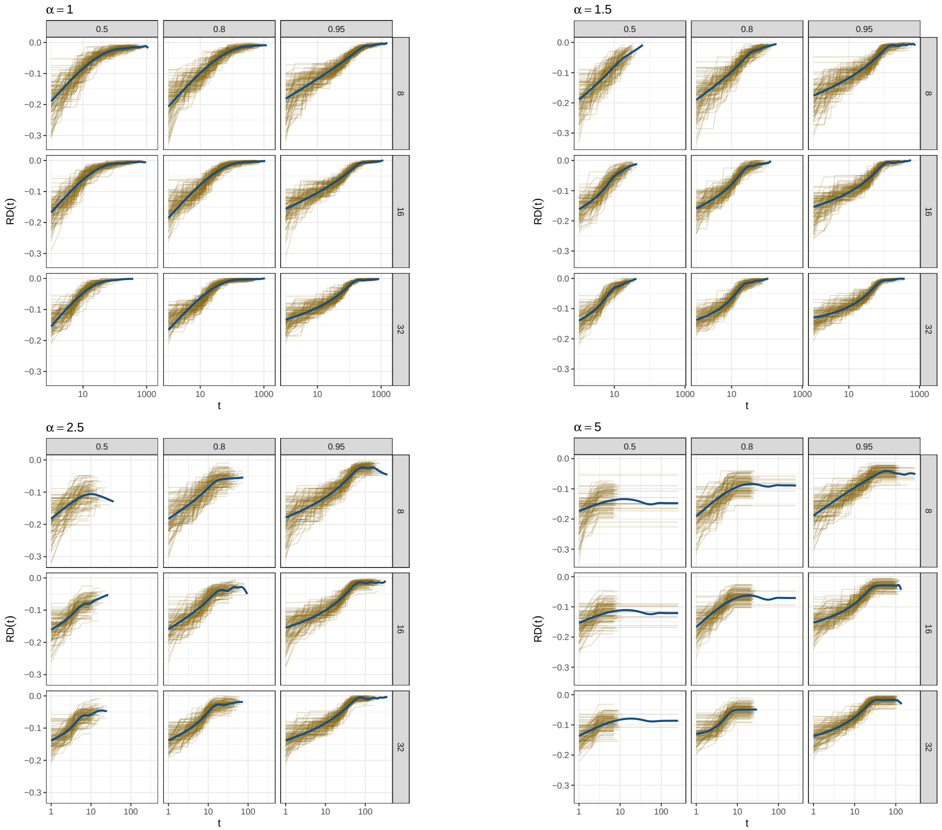


Figure 3.7: The relative deviation from $f(s^{opt})$ as a function of time ($RD[t]$) for the conditions using the iteration-best deposit rule. Yellow lines represent the optimization history for each single replication, blue lines the LOESS-smoothed averages. The x-axis is log-scaled. Each panel shows a different value of α. Within these panels, values of ρ are depicted in columns and values of K are depicted in rows.

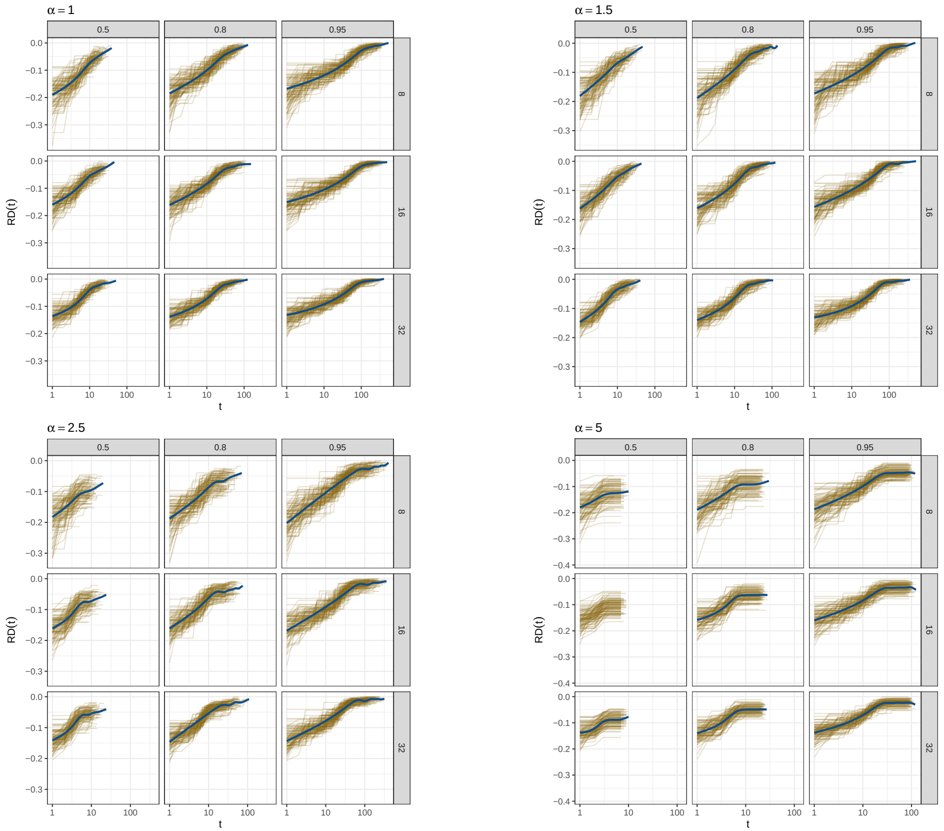


Figure 3.8: The relative deviation from $f(s^{opt})$ as a function of time $(RD[t])$ for the conditions using the global-best deposit rule. Yellow lines represent the optimization history for each single replication, blue lines the LOESS-smoothed averages. The x-axis is log-scaled. Each panel shows a different value of α. Within these panels, values of ρ are depicted in columns and values of K are depicted in rows.

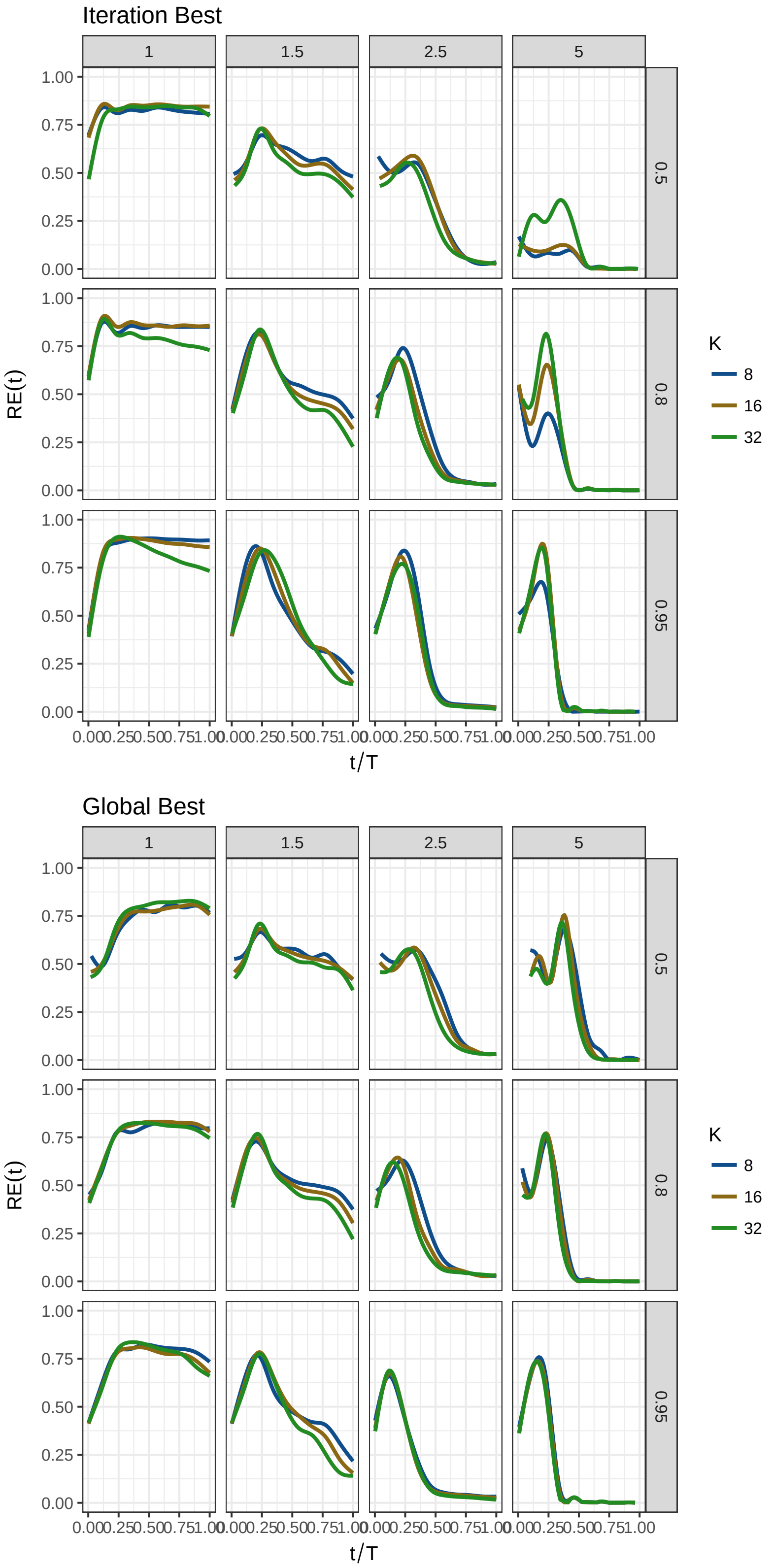


Figure 3.9: Relative Exploration over time. Time is scaled relative to total runtime. Columns depict different settings of α, rows show different settings for ρ.

Table 3.8: Average number of iterations needed for the burn-in phase (T_{bi}) per condition with the average 90th percentile of RE in parentheses.

Deposit	K	ρ	$\alpha = 1$		$\alpha = 1.5$		$\alpha = 2.5$		$\alpha = 5$	
ib	8	0.5	12.303	(0.999)	5.131	(0.834)	2.737	(0.676)	1.990	(0.098)
		0.8	18.465	(0.997)	10.313	(0.895)	5.940	(0.807)	3.323	(0.371)
		0.95	51.400	(0.999)	31.192	(0.930)	18.092	(0.900)	8.000	(0.709)
	16	0.5	17.690	(0.979)	5.250	(0.780)	3.230	(0.666)	2.440	(0.122)
		0.8	27.440	(0.975)	10.760	(0.850)	6.470	(0.742)	3.750	(0.554)
		0.95	76.120	(0.985)	41.640	(0.916)	22.740	(0.845)	11.510	(0.779)
	32	0.5	12.080	(0.926)	4.870	(0.730)	3.170	(0.604)	2.720	(0.326)
		0.8	20.150	(0.906)	12.050	(0.839)	6.700	(0.726)	4.090	(0.679)
		0.95	70.540	(0.936)	45.710	(0.911)	25.630	(0.838)	12.020	(0.763)
gb	8	0.5	7.040	(0.931)	4.371	(0.808)	2.670	(0.667)	2.343	(0.640)
		0.8	15.293	(0.963)	7.454	(0.816)	4.374	(0.718)	2.643	(0.682)
		0.95	38.464	(0.964)	13.707	(0.820)	6.760	(0.689)	5.303	(0.721)
	16	0.5	8.600	(0.898)	4.870	(0.754)	3.090	(0.644)	2.530	(0.587)
		0.8	17.840	(0.931)	8.580	(0.785)	5.520	(0.690)	3.330	(0.663)
		0.95	46.160	(0.899)	26.200	(0.811)	12.480	(0.682)	6.870	(0.687)
	32	0.5	9.260	(0.887)	4.760	(0.719)	3.040	(0.618)	2.700	(0.548)
		0.8	18.960	(0.893)	9.170	(0.765)	5.510	(0.650)	3.520	(0.633)
		0.95	50.440	(0.891)	28.170	(0.784)	13.540	(0.680)	8.440	(0.682)

quality, while also being much slower than all other values of α investigated here. In conditions with $\alpha \in \{2.5, 5\}$ convergence was extremely fast but often lead to poor solutions. The long runtimes for low values of α went hand-in-hand with better - and in some cases sustained - exploration of the space of possible solutions. The comparison of deposit rules becomes relevant only in cases in which good solutions are found slowly (i.e. in situations where $\alpha = 1$). In these cases the *ib* deposit rule outperforms the *gb* with regards to solution quality to a small but noticeable degree, while also being dramatically slower to find solutions. Of all conditions, constellations with $\alpha = 1$, $\rho \in \{.8, .95\}$, $K \in \{16, 32\}$ and the *ib* deposit-rule were able to consistently find the best solutions.

3.7.2 Optimal Heuristic Information

Errors Of the 28800 replications performed for conditions with optimal heuristic information, 10 (0.035%) resulted in errors. As was the case for replications without heuristic information, all of these replications stem from conditions with $K = 8$ and are due to the absence of a viable solution in the first colony.

Success Rate (SR) Over all conditions with optimal heuristics, the average SR was 0.796. Table 3.9 shows the success rates in more detail. All conditions with $\beta = 5$ resulted in recovery

Table 3.9: Condition-specific success rates with optimal heuristic information.

			Iteration-Best				Global-Best			
β	K	ρ	$\alpha = 1$	$\alpha = 1.5$	$\alpha = 2.5$	$\alpha = 5$	$\alpha = 1$	$\alpha = 1.5$	$\alpha = 2.5$	$\alpha = 5$
1	8	0.5	0.810	0.140	0.000	0.000	0.340	0.140	0.000	0.000
		0.8	1.000	0.900	0.140	0.010	0.700	0.650	0.060	0.010
		0.95	1.000	1.000	1.000	0.750	0.960	0.909	0.630	0.220
	16	0.5	0.720	0.400	0.010	0.000	0.480	0.410	0.000	0.000
		0.8	1.000	0.970	0.530	0.040	0.850	0.810	0.240	0.050
		0.95	1.000	1.000	1.000	0.950	0.930	0.910	0.820	0.400
	32	0.5	0.890	0.620	0.040	0.000	0.790	0.660	0.020	0.000
		0.8	1.000	0.980	0.720	0.090	0.920	0.810	0.560	0.070
		0.95	1.000	1.000	1.000	0.970	0.980	0.910	0.960	0.600
1.5	8	0.5	0.910	0.620	0.000	0.000	0.680	0.475	0.020	0.000
		0.8	1.000	0.990	0.710	0.050	0.990	0.940	0.400	0.040
		0.95	1.000	1.000	1.000	1.000	1.000	0.990	0.960	0.780
	16	0.5	0.990	0.860	0.070	0.000	0.890	0.800	0.060	0.000
		0.8	1.000	1.000	0.880	0.340	0.990	0.960	0.720	0.180
		0.95	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.910
	32	0.5	0.980	0.920	0.070	0.010	0.900	0.880	0.280	0.000
		0.8	1.000	1.000	0.980	0.450	0.990	1.000	0.930	0.420
		0.95	1.000	1.000	1.000	1.000	1.000	1.000	0.990	0.970
2.5	8	0.5	1.000	0.990	0.440	0.030	0.990	0.940	0.580	0.040
		0.8	1.000	1.000	1.000	0.810	1.000	1.000	1.000	0.800
		0.95	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	16	0.5	1.000	1.000	0.700	0.080	0.990	1.000	0.820	0.130
		0.8	1.000	1.000	1.000	0.960	1.000	1.000	1.000	1.000
		0.95	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	32	0.5	1.000	0.990	0.920	0.300	1.000	1.000	0.960	0.300
		0.8	1.000	1.000	1.000	1.000	1.000	1.000	0.990	1.000
		0.95	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
5	8	0.5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.95	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	16	0.5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.95	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	32	0.5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0.95	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

of the optimal solution. Generally SR, was extremely high when $\alpha < \beta$ (SR =0.993), especially when compared to those conditions with $\alpha > \beta$ (SR =0.517). When $\alpha \geq \beta$, ρ and K played large roles in determining SR, allowing for sub-optimal solutions to be found only in conditions with very quick evaporation (i.e. when ρ and K are small), however these effects became less pronounced the closer the values of α and β were. Generally, conditions with $\rho = .95$ showed extremely high SR when coupled with the *ib* deposit-rule. Further differences between the deposit rules were visible - albeit small - only when $\beta = 1$ (Cohen's $d_{\beta=1} = 0.276$, 95% $CI = [0.23; 0.322]$). Across all other conditions there was no noticeable difference between the iteration-best and the global-best deposit rules (Cohen's $d_{\beta\neq1} = 0.038$, 95% $CI = [0.011; 0.065]$).

Relative Deviation (RD) The average RD across all conditions with optimal heuristics was -0.006, indicating an average fitness of the final solutions that is less than 1% worse than that of s^{opt}. Figure 3.10 shows the boxplots of the RD for all conditions with optimal heuristics. As was the case for SR, conditions with $\alpha < \beta$ (panels in the lower triangles of Figure 3.10) perform extremely close to optimal, with the average $RD > -1 \times 10^{-6}$ for these conditions.

For those cases in which $\alpha > \beta$, higher values of both the evaporation coefficient ρ and the number of ants K lead to better solution quality. When $\rho = .95$, using the iteration-best deposit rule leads to an $RD > -.0001$, while the global-best deposit rule resulted in $RD = -0.002$. As was the case for SR, those conditions with low values for ρ and larger differences between α and β allowed for quick evaporation and convergence to final solutions much worse than the solution provided in the heuristics. Across all conditions with $\alpha > \beta$ the difference in relative deviation between the two deposit rules was negligible, however (Cohen's $d_{\beta\neq1} = 0.014$, 95% $CI = [-0.023; 0.052]$).

Selection Consistency (SC) The average $SC = 0.958$ across all conditions. Because the SC and RD are very closely related for conditions with optimal heuristics, the same patterns emerge: conditions with $\beta > \alpha$ almost exclusively construct the heuristically favored, optimal solution (in 99.278% of replications) and conditions with larger α than β and low values of ρ construct solutions that are inconsistent with the heuristics provided.

Absolute Runtime (AR) On average, conditions with optimal heuristics required the estimation of 685.376 CFAs, before reaching either of the abort criteria. Figure 3.11 shows the boxplots of the absolute runtimes per condition. Overall, the most important factor for AR was the evaporation coefficient ρ ($\Delta R^2 = 0.672$, 95% $CI = [0.664; 0.681]$), followed by the number of ants ($\Delta R^2 = 0.304$, 95% $CI = [0.297; 0.311]$). In both cases larger values lead to longer absolute runtimes. The opposite is true for the non-linearity coefficient β, where larger values lead to quicker convergence ($\Delta R^2 = 0.135$, 95% $CI = [0.131; 0.139]$).

In total, 6532 of 28800 (22.681%) replications ended after the absolute minimum of iterations

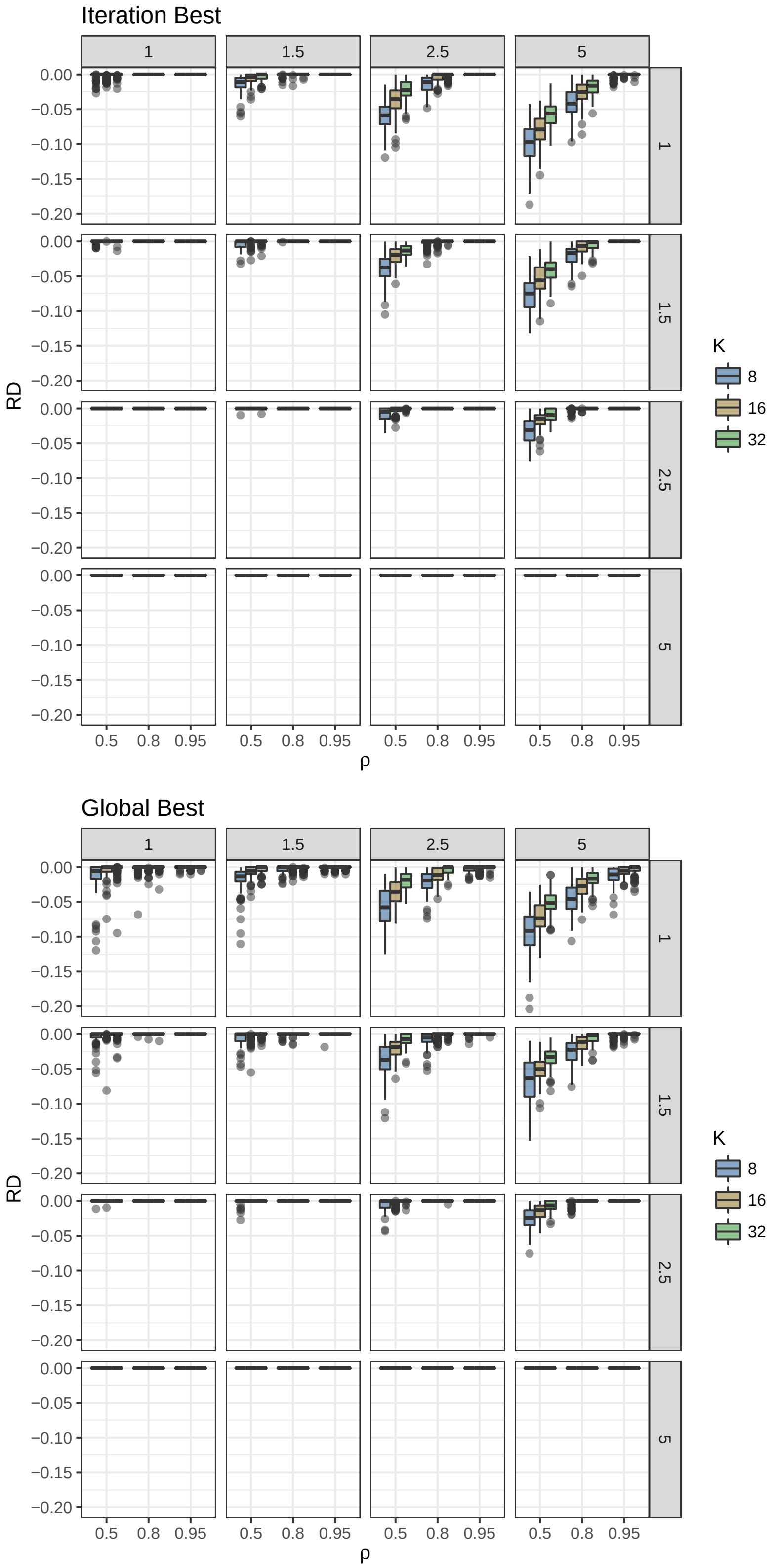
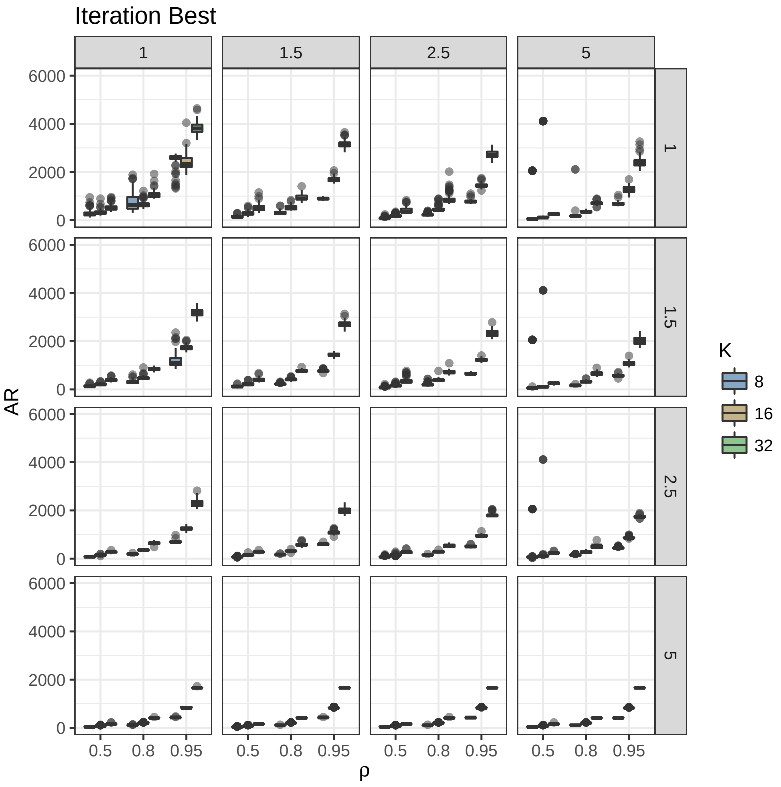


Figure 3.10: Boxplots of the relative deviation (RD) of the final solutions from the optimal solution. Panels represent different settings for α in columns and β in rows.



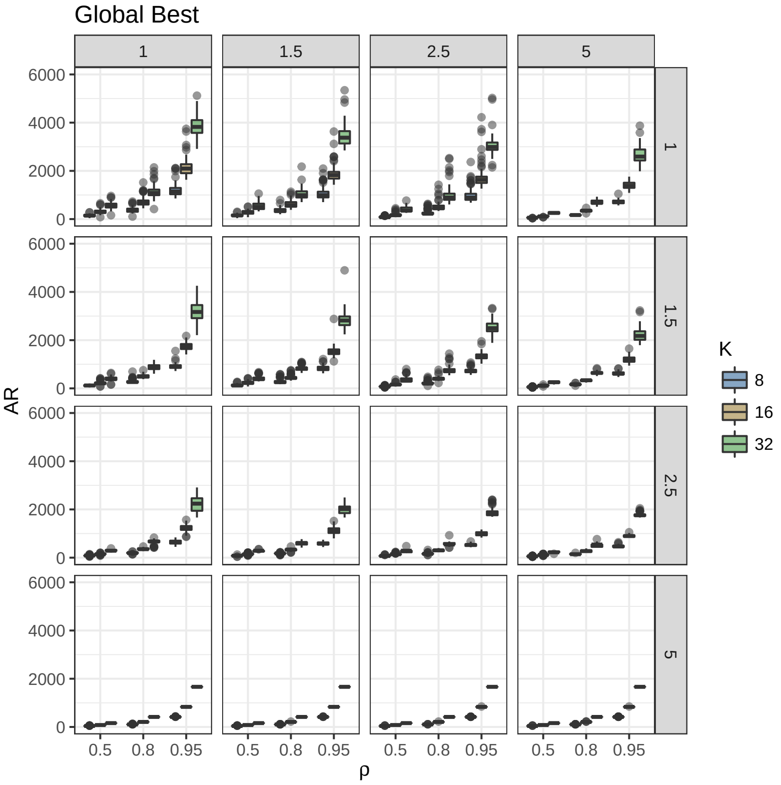


Figure 3.11: Boxplots of the absolute runtime (AR) for all conditions with optimal heuristics. Panels represent different values for α in columns and β in rows.

as derived via Equation (3.9). Of these just 106 replications stemmed from conditions that used values other than $\beta = 5$. Combining $\beta = 5$ with $K = 32$ and the global-best deposit rule lead to minimal AR 100% of the time. With the *ib* deposit rule minimal AR can be seen in 99.5% of those conditions with $\beta = 5$ and $K = 32$. As pointed out in Section 3.6 the probability of constructing the heuristically favored solution in the first iteration is $p = 0.993$ for these cases. When using the *gb* deposit rule this is equivalent to the probability of achieving min AR because the heuristics provided here are optimal. In line with this, smaller values of K (while maintaining $\beta = 5$) lead to less minimal AR replications when using the *ib* deposit to a much greater extent than for those utilizing the *gb* deposit.

Relative Exploration (RE)

Across all conditions the average $RE = 0.382$, indicating that just over a third of all constructed solutions were unique and viable solutions. Figure 3.12 shows the boxplots of the RE for all conditions. In contrast to the results regarding AR, the parameter with the most influence on relative exploration was α. This is indicated by the downward shift across the panels of Figure 3.12 from left to right, meaning that higher values of α lead to less exploration overall. Additionally, the effect size ΔR^2 attributable to α and all its interactions is $\Delta R^2 = 0.738$ (95% $CI = [0.729; 0.746]$) - making it, by far, the most the most important determinant of RE among the parameters varied in this study. To illustrate: the 53 conditions with the highest RE values all utilized $\alpha = 1$.

The parameter with the second largest relevance for RE was the non-linearity coefficient β with $\Delta R^2 = 0.188$ (95% $CI = [0.182; 0.193]$). As was the case for α, higher values of β lead to less exploration. The pronounced effects of the relation between α and β, seen with regards to solutions quality, did not appear for the relative exploration. The same trend holds true for the evaporation coefficient ρ ($\Delta R^2_\rho = 0.124$, 95% $CI = [0.12; 0.128]$). The deposit rule did not have an impact on exploration for situations with optimal heuristics (Cohen's $d = 0.017$, 95% $CI = [-0.006; 0.04]$).

Relative Deviation over Time ($RD[t]$) Figures 3.13 through 3.16 depict the $RD(t)$ of all replications using the optimal heuristics that did not result in errors. Yellow lines are the $RD(t)$ of single replications and the blue lines represent the LOWESS lines across all replications per condition. The scaling of the x-axes is different across the figures due to the extreme differences in AR between conditions.

For all conditions with $\beta = 5$, the $RD(t)$ quickly moved to 0, taking an average of 1.151 iterations to obtain the final solution. After this short exploration, the s^{gb} was iterated upon an average of 22.363 times before convergence. Conditions with $\beta > \alpha$ exhibit short search processes (M_T =7.549), with a clear trend towards $RD(t) = 0$ and a long phase of replicating the final solution (M_T =24.561). Over all conditions, iterations reiterating on the same s^{gb} make

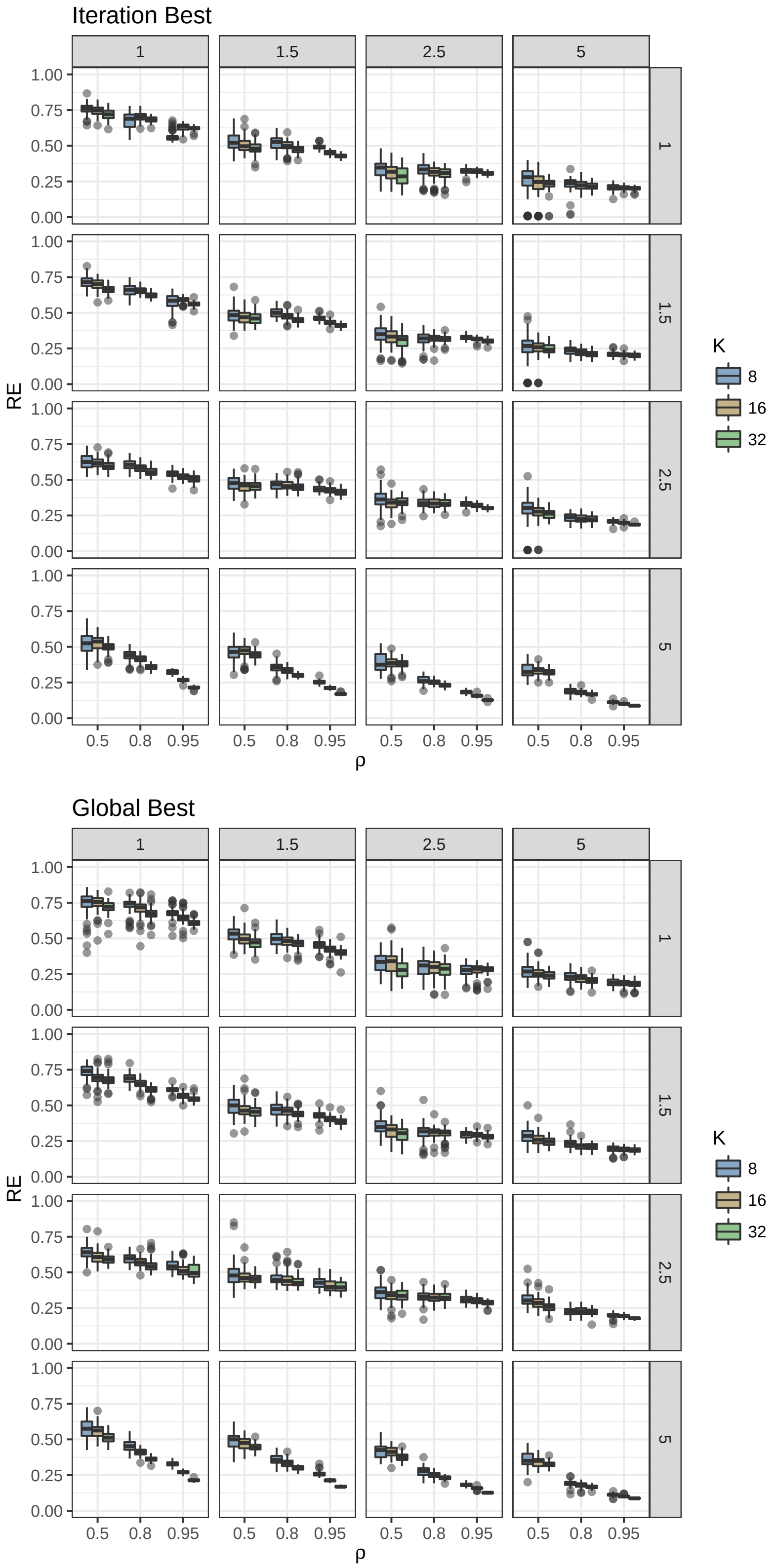


Figure 3.12: Boxplots of the relative exploration (RE) for all conditions with optimal heuristics. Panels represent different values for α in columns and β in rows.

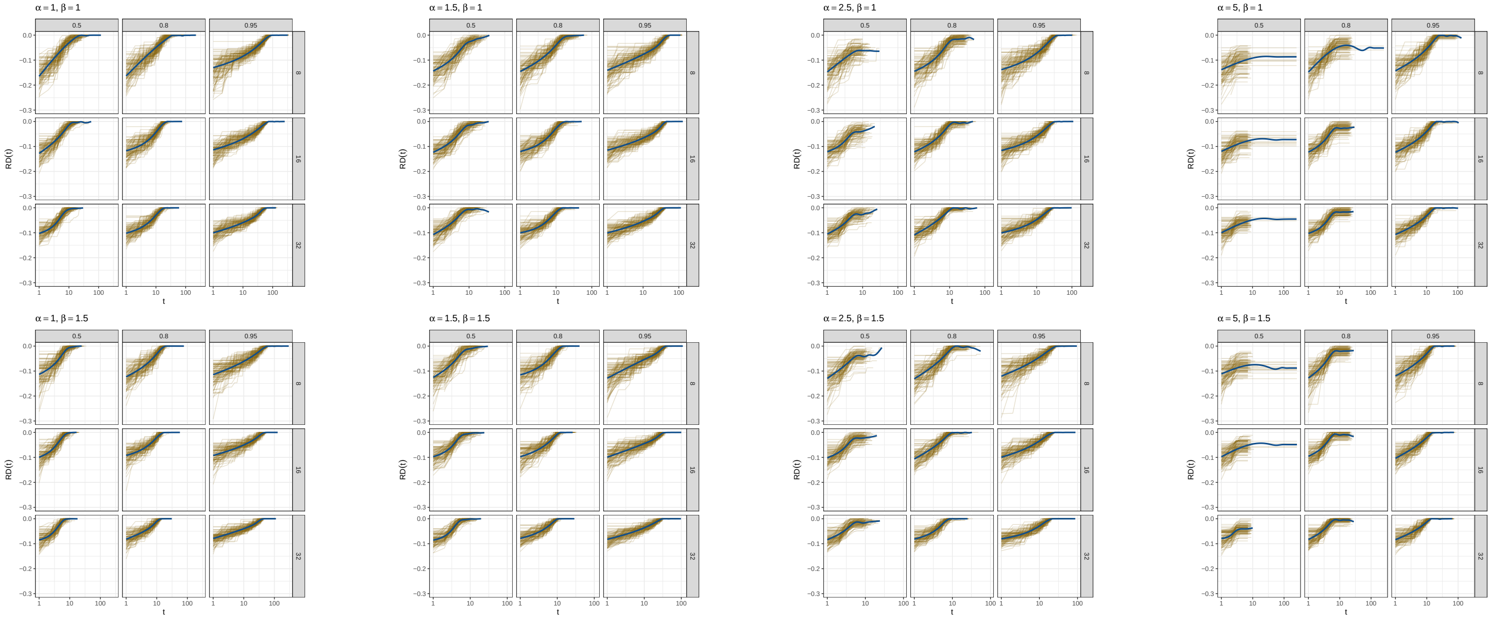


Figure 3.13: The relative deviation from $f(s^{opt})$ as a function of time $(RD[t])$ for the conditions using the iteration-best deposit rule and $\beta = 1$ or $\beta = 1.5$. Yellow lines represent the optimization history for each single replication, blue lines the LOESS-smoothed averages. The x-axis is log-scaled. The columns of each panel show different values of ρ, the rows different values of K.

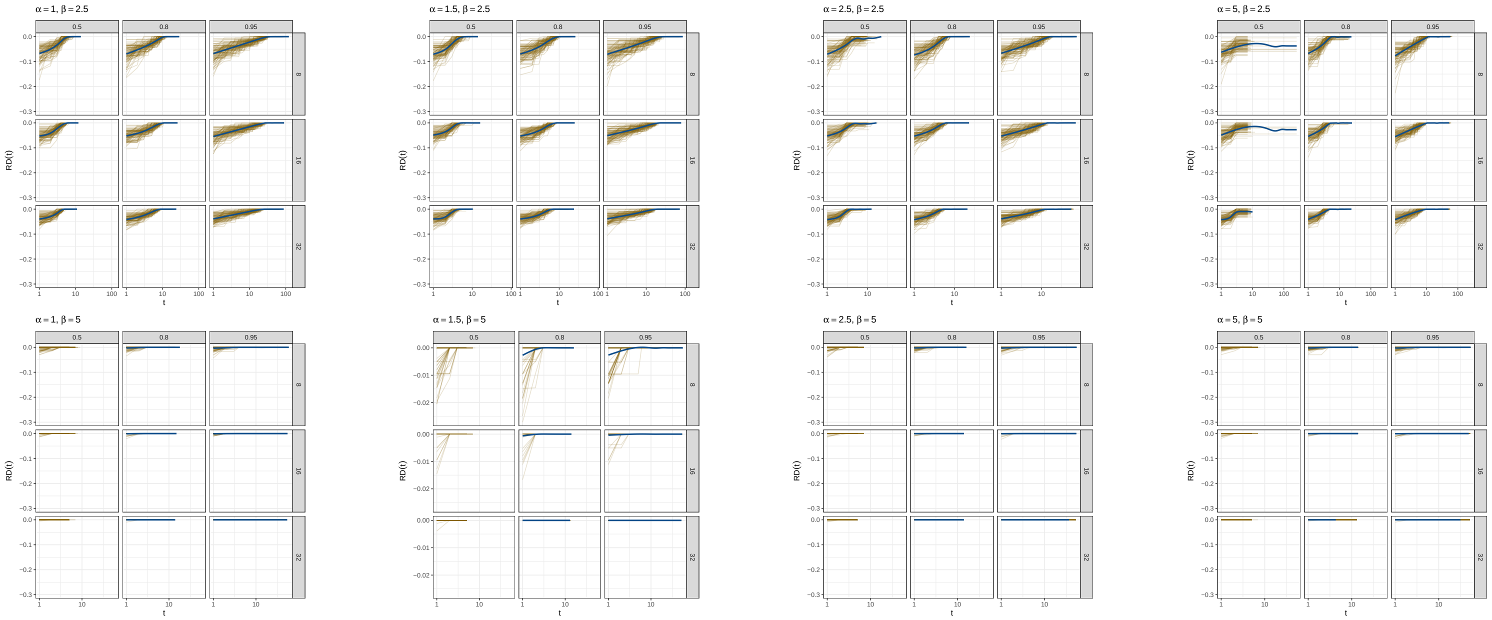


Figure 3.14: The relative deviation from $f(s^{opt})$ as a function of time ($RD[t]$) for the conditions using the iteration-best deposit rule and $\alpha = 2.5$ or $\alpha = 5$. Yellow lines represent the optimization history for each single replication, blue lines the LOESS-smoothed averages. The x-axis is log-scaled. The columns of each panel show different values of ρ, the rows different values of K.

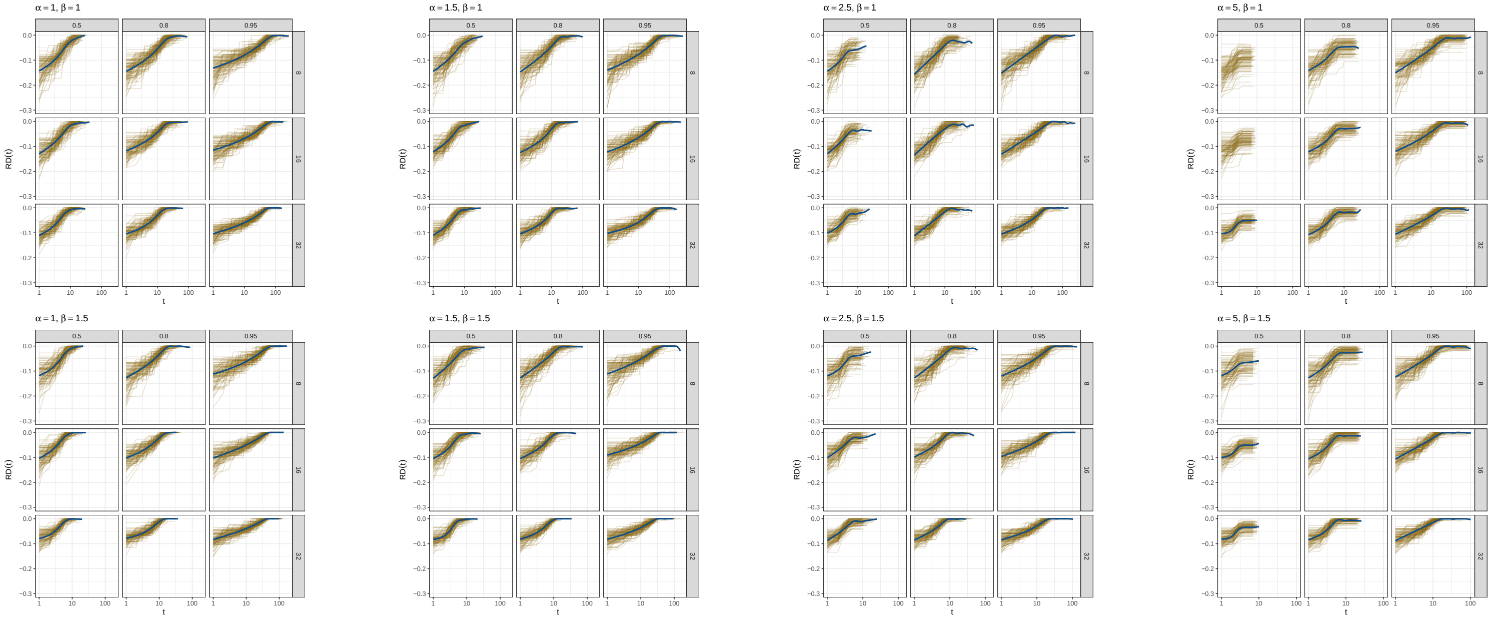


Figure 3.15: The relative deviation from $f(s^{opt})$ as a function of time $(RD[t])$ for the conditions using the global-best deposit rule and $\beta = 1$ or $\beta = 1.5$. Yellow lines represent the optimization history for each single replication, blue lines the LOESS-smoothed averages. The x-axis is log-scaled. The columns of each panel show different values of ρ, the rows different values of K.

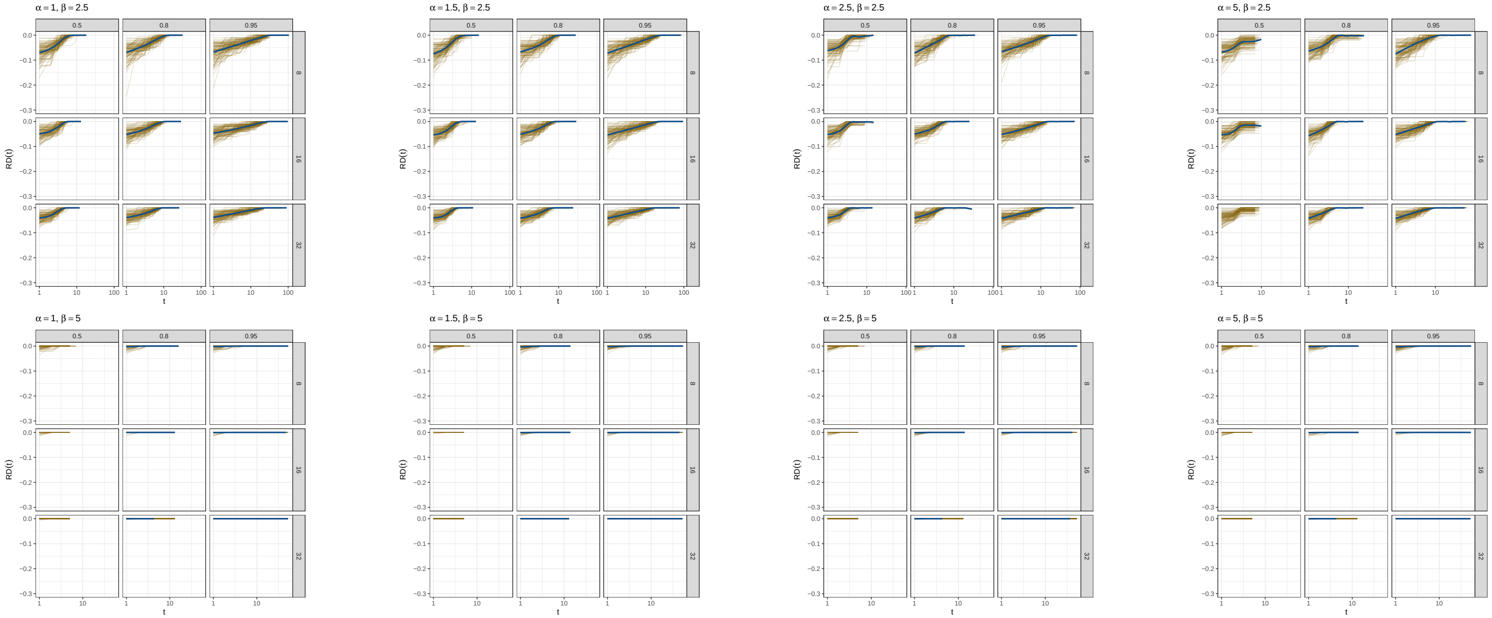


Figure 3.16: The relative deviation from $f(s^{opt})$ as a function of time ($RD[t]$) for the conditions using the global-best deposit rule and $\beta = 2.5$ or $\beta = 5$. Yellow lines represent the optimization history for each single replication, blue lines the LOESS-smoothed averages. The x-axis is log-scaled. The columns of each panel show different values of ρ, the rows different values of K.

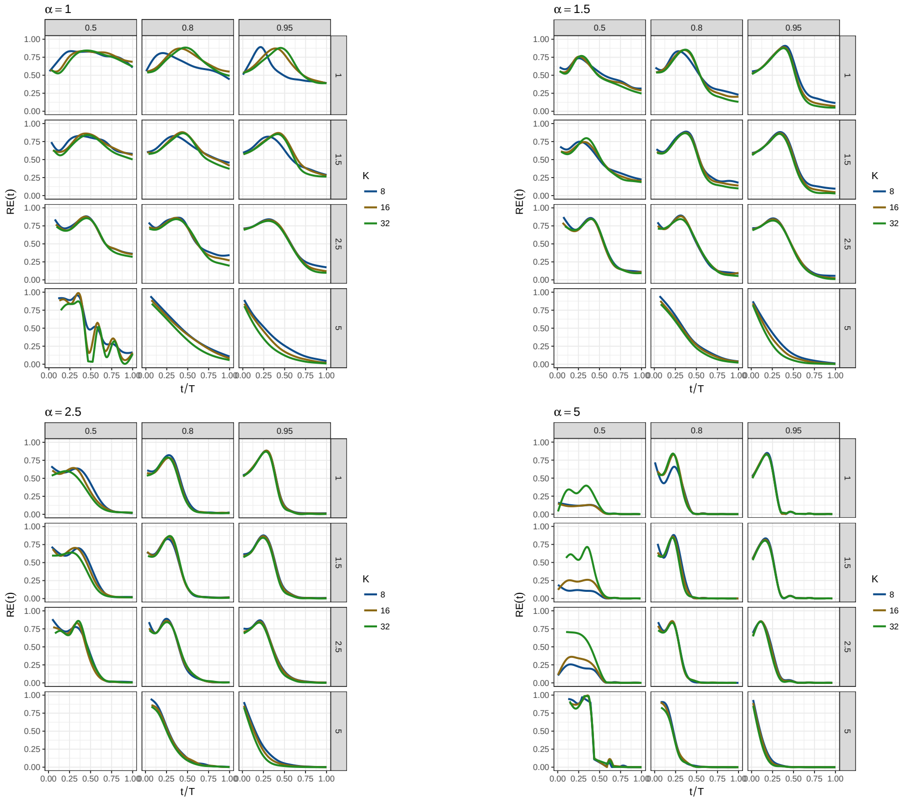


Figure 3.17: The LOWESS lines of the relative exploration over time for conditions with iteration-best deposit rule. Different values of ρ are in columns, different values of β in rows, and different values of α are in different panels.

up roughly 67% of the entire search duration. Though the search for s^{gb} is longer for conditions with $\alpha > \beta$ ($M_T = 13.515$), the RD of in these conditions is often lower than that in those conditions favoring the heuristic information more heavily. In conditions with $\alpha = 5$ and low values for β, very few new best solutions were generated before converging into sub-optimal solutions.

Relative Exploration over Time ($RE[t]$) Figures 3.17 and 3.18 show the timed relative exploration when using optimal heuristics as LOWESS lines. Empty panels are due to information being too sparse for the LOWESS smoother implemented in the `stats`-Package (Version 3.3.2; R Core Team, 2016) to handle. As pointed out above, this is due to conditions with $\beta = 5$ often requiring the absolute minimum of iterations to converge.

In almost all conditions with $\beta \neq 5$, the $RE(t)$ reached its peak after a short burn-in phase before deteriorating quickly. The notable exception were conditions with $\alpha = 5$, $\rho = .5$, and the iteration-best deposit rule, for which $RE(t)$ never achieved acceptable values. Conditions with

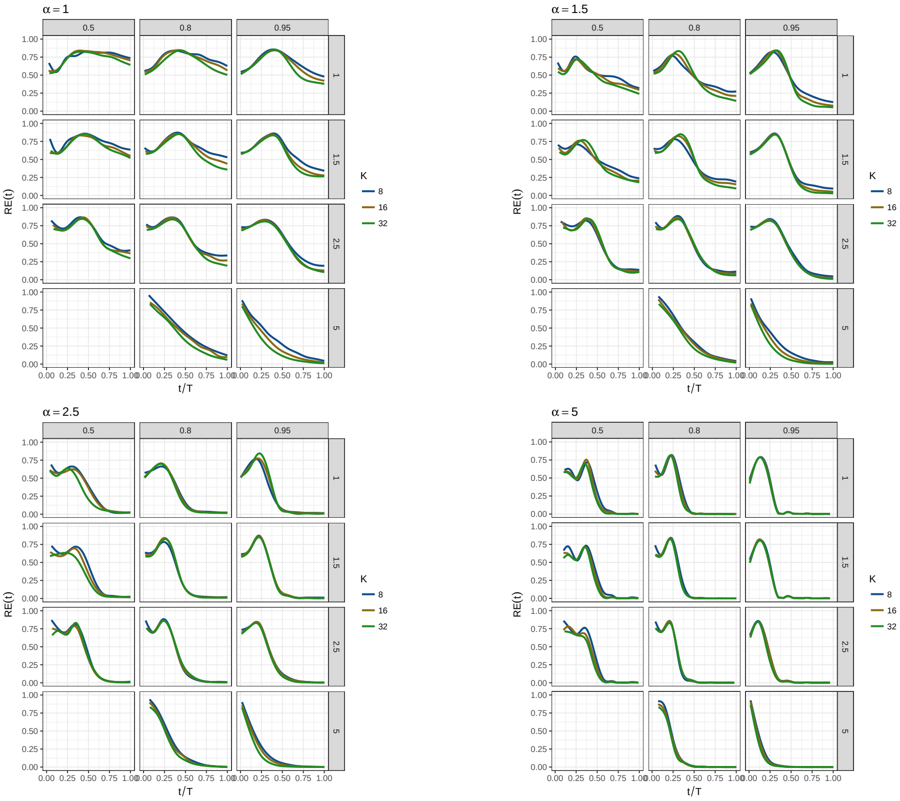


Figure 3.18: The LOWESS lines of the relative exploration over time for conditions with global-best deposit rule. Different values of ρ are in columns, different values of β in rows, and different values of α are in different panels.

$\beta = 5$ almost always started at their maximum $RE(t)$ and showed a quick and steep decline in exploration. Somewhat irrespective of other parameter settings, larger values of α lead to a shorter period of initial exploration before a period of reiterating the final solution. In cases with $\alpha \in \{2.5, 5\}$ iterations with an $RE(t) = 0$ constitute 51.626% of all iterations. In the worst case - when $\alpha = 5$, $\beta = 1.5$, $\rho = 0.5$, $K = 8$, and using the *ib* deposit rule - iterations with $RE(t) = 0$ made up 90.351% of all iterations. The most consistent exploration was achieved in conditions utilizing a linear pheromone deposit as well as $\beta = 1$. As pointed out above, these cases were also those with the highest RE overall.

Tables 3.10 and 3.11 depict the 90th percentile in $RE(t)$ and the average number of iterations required to reach this $RE(t)$. Conditions with large values for β reach their 90th percentile $RE(t)$ very quickly - sometimes beginning at their maximum exploration - but tend to have a lower overall peak performance. The highest 90th percentile $RE(t)$ was reached in the condition with $\rho = 0.5$, $K = 8$, $\alpha = 1$, $\beta = 1$ and utilizing the *ib* deposit rule at a value of 0.959. Overall, the 14 conditions with the highest 90th percentile $RE(t)$ stem from conditions with $\alpha = 1$ and the 12 best performing conditions used $K = 8$. Among these conditions, those with larger values of β and lower values of ρ tended to surmount the initial burn-in quickest.

Summary For conditions with optimal heuristics, there is a general tendency for quick and reliable recovery of the optimal solution in cases in which $\beta > \alpha$. Exploration is dependent mostly upon the values of α with a clear tendency for lower values of α to lead to more exploration. Overall, only conditions with large values of α, small values of ρ, and small values of β performed poorly in recovery of the optimal solution and showed little exploration. In conditions in which β was large the search showed little exploration, discovering the optimal solution very early on and iterating on this final solution for most of the runtime. Over all conditions with optimal heuristics, the iteration-best deposit rule tended to show better performance in recovering the optimal solution.

3.7.3 Sub-Optimal Heuristic Information

Errors 64 of the 28800 (0.222%) replications using the median solution to provide heuristic information resulted in errors. As was the case in the previous two sections, all of these errors were due to the failure to find an admissible solution in the first colony. All of the errors occurred in conditions with $K = 8$.

Success Rate Table 3.12 provides the success rates for conditions implementing the iteration-best and the global-best deposit rule, respectively. Overall, the average $SR = 0.007$, indicating a very low rate of recovering the optimal solution. Specifically, all conditions with $\beta \geq 2.5$ showed a $SR = 0$ and all but one condition with $\alpha \geq 2.5$ also never recovered the optimal solution. The

Table 3.10: Average number of iterations needed for the burn-in phase (T_{bi}) per condition utilizing the iteration-best deposit rule. The average 90th percentile of RE is in parentheses.

β	K	ρ	$\alpha = 1$		$\alpha = 1.5$		$\alpha = 2.5$		$\alpha = 5$	
1	8	0.5	6.750	(0.959)	3.670	(0.819)	2.740	(0.764)	1.690	(0.147)
		0.8	8.270	(0.914)	6.650	(0.895)	4.570	(0.858)	2.606	(0.624)
		0.95	12.293	(0.877)	13.030	(0.908)	8.828	(0.864)	5.040	(0.806)
	16	0.5	6.890	(0.910)	4.050	(0.778)	2.670	(0.701)	2.260	(0.135)
		0.8	11.930	(0.905)	8.390	(0.872)	5.470	(0.801)	3.230	(0.724)
		0.95	29.450	(0.892)	21.550	(0.872)	12.320	(0.840)	6.030	(0.769)
	32	0.5	6.170	(0.876)	4.140	(0.759)	3.030	(0.681)	2.440	(0.363)
		0.8	12.520	(0.887)	9.310	(0.852)	5.830	(0.795)	3.280	(0.720)
		0.95	35.790	(0.868)	25.500	(0.850)	15.450	(0.819)	7.720	(0.753)
1.5	8	0.5	4.560	(0.932)	2.830	(0.824)	2.570	(0.772)	1.616	(0.128)
		0.8	6.750	(0.915)	5.160	(0.895)	3.980	(0.844)	2.150	(0.796)
		0.95	9.750	(0.890)	8.630	(0.889)	5.100	(0.866)	3.300	(0.795)
	16	0.5	5.000	(0.880)	3.590	(0.789)	2.790	(0.751)	2.160	(0.272)
		0.8	9.150	(0.889)	7.180	(0.859)	4.420	(0.824)	2.780	(0.743)
		0.95	20.860	(0.877)	14.240	(0.859)	9.370	(0.827)	4.200	(0.764)
	32	0.5	4.900	(0.845)	3.720	(0.790)	2.820	(0.723)	2.340	(0.638)
		0.8	9.100	(0.852)	6.970	(0.838)	5.140	(0.805)	3.120	(0.719)
		0.95	24.970	(0.844)	17.440	(0.827)	11.300	(0.805)	5.600	(0.744)
2.5	8	0.5	2.520	(0.911)	2.140	(0.876)	2.030	(0.841)	1.280	(0.244)
		0.8	3.100	(0.920)	2.550	(0.902)	2.030	(0.864)	1.680	(0.818)
		0.95	3.410	(0.893)	3.230	(0.884)	2.270	(0.871)	1.530	(0.796)
	16	0.5	3.350	(0.865)	2.890	(0.829)	2.460	(0.796)	1.500	(0.352)
		0.8	4.250	(0.873)	3.740	(0.864)	2.790	(0.822)	1.940	(0.788)
		0.95	6.240	(0.865)	4.870	(0.854)	3.420	(0.827)	2.000	(0.772)
	32	0.5	3.620	(0.832)	3.090	(0.807)	2.620	(0.775)	1.450	(0.683)
		0.8	4.950	(0.838)	4.010	(0.825)	3.190	(0.801)	2.300	(0.760)
		0.95	8.350	(0.832)	6.750	(0.823)	4.280	(0.804)	1.980	(0.751)
5	8	0.5	1.200	(0.891)	1.143	(0.878)	1.190	(0.896)	1.140	(0.895)
		0.8	1.110	(0.834)	1.040	(0.825)	1.020	(0.802)	1.030	(0.733)
		0.95	1.040	(0.749)	1.020	(0.693)	1.000	(0.617)	1.000	(0.445)
	16	0.5	1.210	(0.859)	1.120	(0.845)	1.180	(0.845)	1.180	(0.831)
		0.8	1.050	(0.788)	1.030	(0.765)	1.030	(0.745)	1.000	(0.698)
		0.95	1.020	(0.667)	1.000	(0.631)	1.000	(0.556)	1.000	(0.413)
	32	0.5	1.130	(0.789)	1.230	(0.801)	1.190	(0.800)	1.160	(0.795)
		0.8	1.020	(0.733)	1.010	(0.717)	1.000	(0.703)	1.000	(0.659)
		0.95	1.000	(0.589)	1.000	(0.559)	1.000	(0.484)	1.000	(0.338)

Table 3.11: Average number of iterations needed for the burn-in phase (T_{bi}) per condition utilizing the global-best deposit rule. The average 90th percentile of RE is in parentheses.

β	K	ρ	$\alpha = 1$		$\alpha = 1.5$		$\alpha = 2.5$		$\alpha = 5$	
1	8	0.5	6.170	(0.946)	3.530	(0.803)	2.450	(0.738)	1.909	(0.663)
		0.8	10.170	(0.955)	5.630	(0.835)	3.800	(0.775)	2.140	(0.738)
		0.95	18.667	(0.939)	10.354	(0.881)	7.120	(0.805)	3.280	(0.757)
	16	0.5	7.060	(0.910)	3.590	(0.744)	2.860	(0.707)	2.370	(0.636)
		0.8	11.790	(0.902)	7.360	(0.830)	4.710	(0.764)	3.130	(0.718)
		0.95	27.000	(0.882)	18.540	(0.851)	10.410	(0.791)	4.680	(0.733)
	32	0.5	6.450	(0.875)	3.830	(0.735)	2.950	(0.670)	2.480	(0.619)
		0.8	11.750	(0.861)	8.320	(0.835)	5.060	(0.745)	3.330	(0.687)
		0.95	32.250	(0.857)	22.780	(0.824)	13.480	(0.796)	6.530	(0.712)
1.5	8	0.5	4.460	(0.940)	2.646	(0.812)	2.360	(0.781)	1.780	(0.711)
		0.8	7.260	(0.933)	4.460	(0.858)	3.320	(0.815)	1.990	(0.780)
		0.95	10.080	(0.899)	6.040	(0.873)	4.570	(0.847)	2.890	(0.770)
	16	0.5	4.870	(0.878)	3.190	(0.794)	2.870	(0.742)	2.090	(0.660)
		0.8	8.300	(0.879)	6.280	(0.844)	4.300	(0.817)	2.760	(0.734)
		0.95	16.760	(0.863)	12.540	(0.845)	7.770	(0.815)	4.150	(0.752)
	32	0.5	5.030	(0.855)	3.730	(0.777)	2.900	(0.710)	2.240	(0.633)
		0.8	9.120	(0.845)	6.910	(0.820)	4.750	(0.801)	2.920	(0.710)
		0.95	21.690	(0.833)	15.800	(0.819)	10.270	(0.792)	5.300	(0.733)
2.5	8	0.5	2.630	(0.906)	2.200	(0.878)	2.050	(0.840)	1.440	(0.784)
		0.8	3.180	(0.924)	2.350	(0.894)	2.180	(0.871)	1.430	(0.808)
		0.95	3.600	(0.890)	2.910	(0.877)	2.120	(0.867)	1.480	(0.779)
	16	0.5	3.410	(0.855)	2.940	(0.844)	2.370	(0.790)	1.340	(0.727)
		0.8	3.790	(0.874)	3.480	(0.854)	2.650	(0.831)	1.830	(0.777)
		0.95	5.320	(0.859)	4.800	(0.849)	3.460	(0.822)	1.860	(0.761)
	32	0.5	3.590	(0.829)	3.100	(0.808)	2.670	(0.762)	1.340	(0.683)
		0.8	5.090	(0.840)	3.760	(0.824)	3.040	(0.808)	2.010	(0.757)
		0.95	7.230	(0.827)	5.940	(0.820)	4.310	(0.801)	2.070	(0.740)
5	8	0.5	1.110	(0.904)	1.120	(0.904)	1.090	(0.900)	1.090	(0.885)
		0.8	1.000	(0.857)	1.030	(0.839)	1.020	(0.795)	1.000	(0.743)
		0.95	1.040	(0.740)	1.020	(0.697)	1.000	(0.621)	1.000	(0.460)
	16	0.5	1.180	(0.848)	1.150	(0.841)	1.120	(0.867)	1.070	(0.842)
		0.8	1.090	(0.778)	1.030	(0.768)	1.000	(0.739)	1.000	(0.685)
		0.95	1.000	(0.675)	1.000	(0.633)	1.000	(0.560)	1.000	(0.386)
	32	0.5	1.130	(0.806)	1.180	(0.796)	1.210	(0.808)	1.140	(0.798)
		0.8	1.050	(0.732)	1.030	(0.721)	1.000	(0.708)	1.000	(0.662)
		0.95	1.000	(0.604)	1.000	(0.557)	1.000	(0.474)	1.000	(0.312)

Table 3.12: Condition-specific success rates (SR) with sub-optimal heuristic information.

			Iteration-Best				Global-Best			
β	K	ρ	$\alpha = 1$	$\alpha = 1.5$	$\alpha = 2.5$	$\alpha = 5$	$\alpha = 1$	$\alpha = 1.5$	$\alpha = 2.5$	$\alpha = 5$
1	8	0.5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.95	0.000	0.000	0.000	0.000	0.020	0.112	0.000	0.000
	16	0.5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.8	0.000	0.020	0.000	0.000	0.000	0.030	0.000	0.000
		0.95	0.000	0.030	0.000	0.000	0.110	0.160	0.000	0.000
	32	0.5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.8	0.110	0.020	0.000	0.000	0.000	0.050	0.000	0.000
		0.95	0.000	0.110	0.000	0.000	0.180	0.310	0.020	0.000
1.5	8	0.5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.8	0.000	0.000	0.000	0.000	0.000	0.010	0.000	0.000
		0.95	0.000	0.000	0.000	0.000	0.000	0.030	0.000	0.000
	16	0.5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.95	0.000	0.000	0.000	0.000	0.040	0.110	0.000	0.000
	32	0.5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.8	0.000	0.020	0.000	0.000	0.010	0.010	0.000	0.000
		0.95	0.000	0.050	0.000	0.000	0.110	0.190	0.000	0.000
2.5	8	0.5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.95	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	16	0.5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.95	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	32	0.5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.95	0.000	0.000	0.000	0.000	0.000	0.020	0.000	0.000
5	8	0.5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.95	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	16	0.5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.95	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	32	0.5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.95	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

one exception is the condition with $\alpha = 2.5$, $\beta = 1$, $\rho = .95$, $K = 32$, and the global-best deposit rule, which found the optimum twice in 100 replications.

Among the remaining 36 conditions with $\alpha \in \{1, 1.5\}$ and $\beta \in \{1, 1.5\}$, using the iteration-best deposit rule showed slightly lower success rates (Cohen's $d = -0.198$, 95% $CI = [-0.245; -0.152]$). Additionally, conditions with $\alpha = 1.5$ performed marginally better than those with $\alpha = 1$ (Cohen's $d = 0.12$, 95% $CI = [0.074; 0.167]$). Across all conditions the best performance ($SR = 0.31$) was achieved with $\alpha = 1.5$, $\beta = 1$, $\rho = .95$, $K = 32$, and the global-best deposit rule.

Relative Deviation (RD) Across all conditions, the mean RD = -0.074 meaning that on average, the final solution performed approximately 7% worse than the optimal solution. Keep in mind that in all of the conditions discussed in this section, the heuristic information favored the median solution, which performs to an RD = -0.238.

Figure 3.19 shows the boxplots of the RD for all conditions with sub-optimal heuristics. Higher values of β are accompanied by worse performance with regards to RD, as visualized by the downward shift in the different panel-rows of Figure 3.19. Numerically, this can be shown via the RD means for the different levels of β, with $RD_{\beta=1} = -0.044$, $RD_{\beta=1.5} = -0.052$, $RD_{\beta=2.5} = -0.073$, and $RD_{\beta=5} = -0.126$. Overall, the levels of β and all its interactions account for $\Delta R^2 = 0.524$ (95% $CI = [0.515; 0.534]$) in the total variation of RD among conditions with sub-optimal heuristics.

The second most prominent determinant of RD in these conditions is the non-linearity coefficient α, which accounts for $\Delta R^2 = 0.176$ (95% $CI = [0.169; 0.182]$) with its main-effect and all its interactions. In line with previous findings, higher values of α again lead to worse values in RD. The evaporation coefficient ρ also showed consistent effects with larger values of ρ leading to solutions closer to s^{opt} in quality ($\Delta R^2 = 0.123$, 95% $CI = [0.117; 0.129]$). Similarly, an increase in the number of ants lead to RD values closer to 0 ($\Delta R^2 = 0.076$, 95% $CI = [0.07; 0.081]$).

Of all manipulated factors the deposit rule had the least impact on RD ($\Delta R^2 = 0.027$, 95% $CI = [0.023; 0.032]$) and also leads to the least clear results. In approximately 81% of all paired conditions the global-best deposit rule outperformed the iteration-best. Of the remaining 28 conditions, only 6 resulted in a difference in RD of more than .01. All of these conditions used $\alpha = 1$ and $\rho = .5$.

Table 3.13 shows the four best performing conditions with respect to average and worst-case RD, which differ only in deposit rule and α being either 1 or 1.5. The worst solution constructed in these conditions resulted in $f(s) = 1.895$, which stems from a shortened scale with $crel = 0.925$ and a measurement model with good fit (RMSEA = 0.029, SRMR = 0.027, CFI = 0.969).

When computing RD relative to the median solution which served as the foundation of the heuristic information used in these conditions, the average $RD^{md} = 0.216$, indicating an average performance of the final solutions that was considerably better than the heuristically favored solution. Over all 28736 replications which resulted in admissible solutions, only 1

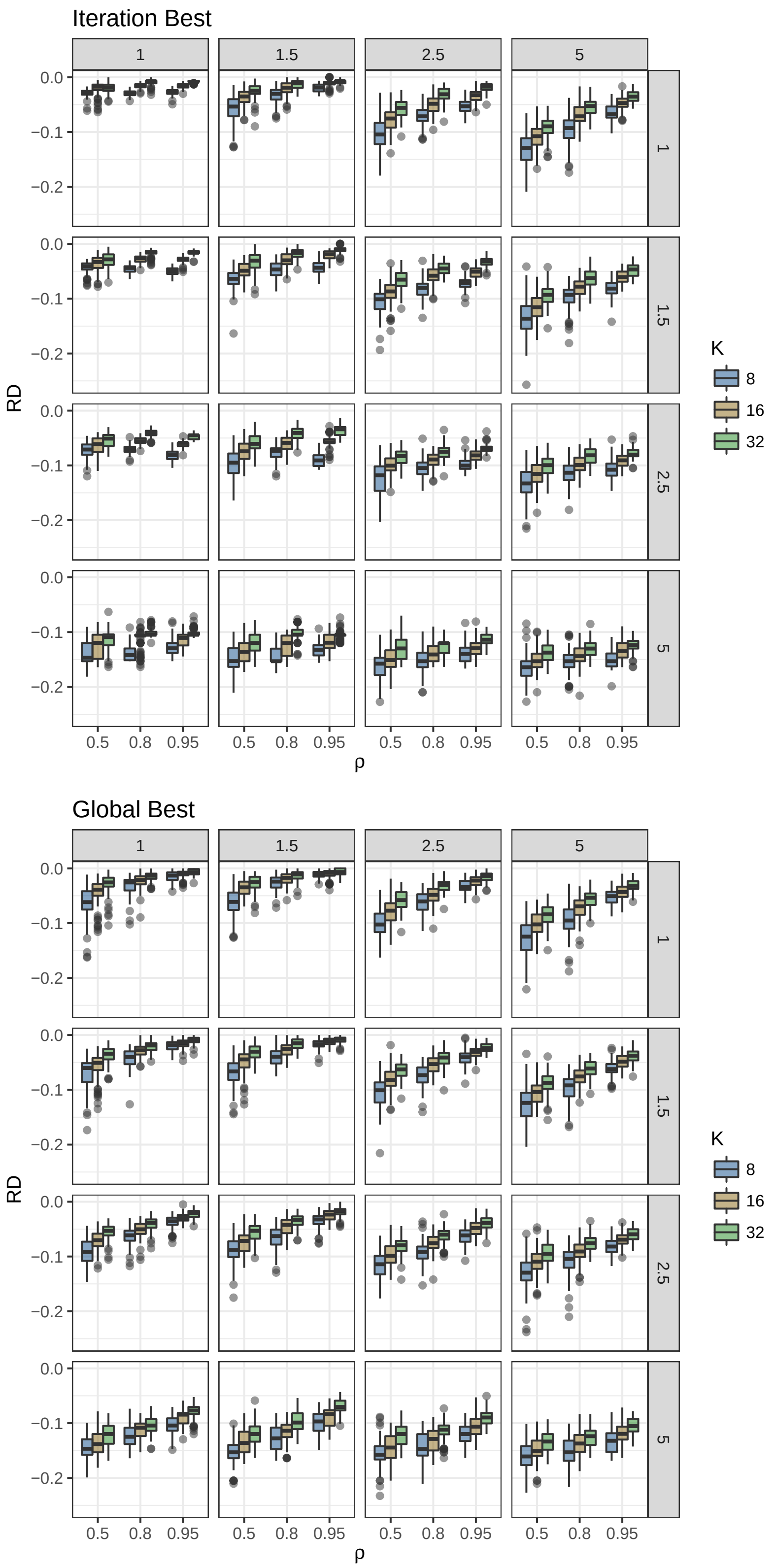


Figure 3.19: Boxplots of the relative deviation (RD) of the final solutions from the optimal solution. Panels represent different values for α in columns and β in rows.

Table 3.13: Average and worst-case *RD* for the four best performing conditions. The ranks pertain to the total of all 288 conditions without heuristic information.

					Average		Worst-Case			
Deposit	α	β	K	ρ	RD	Rank	RD	Rank	SR	AR
ib	1	1	32	0.95	-0.008	4	-0.013	1	0.000	21876.480
ib	1.5	1	32	0.95	-0.008	3	-0.021	2	0.000	23840.320
gb	1	1	32	0.95	-0.007	2	-0.027	3	0.180	8931.200
gb	1.5	1	32	0.95	-0.007	1	-0.027	4	0.110	9680.640

showed an $RD^{md} < 0$. All remaining replications resulted in solutions that outperformed s^{md}. As described in Section 3.5, the RD^{md} is simply a linear transformation of the RD with relation to optimal solution and therefore the comparison of different conditions with regards to RD^{md} is not reported here.

Selection Consistency (SC) For the conditions implementing the sub-optimal heuristics the average $SC = 0.667$, indicating that roughly 12 (out of a total of 18) items are shared between the final solutions and the median solution. As pointed out in Section 3.4, the optimal and the median solution share 7 common items, resulting in an $SC = 0.389$. Figure 3.20 shows the boxplots of the SC for all conditions implementing the sub-optimal heuristics.

The non-linearity coefficient β was, by far, the most important determinant of selection consistency, with its main effect and all its interactions accounting for a $\Delta R^2 = 0.731$ (95% $CI = [0.724; 0.739]$). Generally, higher values of β lead to solutions that resemble the heuristically favored solution more, with the average SC for conditions with $\beta = 5$ at 0.871. Overall, the highest condition-specific average SC was 0.908, encountered in the condition with $\alpha = 2.5$, $\beta = 2.5$, $\rho = 0.5$, $K = 8$, and the *gb* deposit rule. It should also be noted, that not a single replication selected exactly the 16 items that were preferred by the heuristics, nor was there any replication which selected none of those 16 items.

The remaining four experimental factors had small effects on the SC. The non-linearity coefficient α achieved a $\Delta R^2 = 0.057$ (95% $CI = [0.051; 0.062]$), while ρ ($\Delta R^2 = 0.044$, 95% $CI = [0.038; 0.049]$), K ($\Delta R^2 = 0.036$, 95% $CI = [0.031; 0.042]$), and the deposit rule ($\Delta R^2 = 0.043$, 95% $CI = [0.038; 0.049]$) all achieve ΔR^2 values below .05. Generally, fewer ants per colony lead to higher consistency. For α and ρ lower values lead to solutions more consistent with the heuristics. Across all conditions, the iteration-best deposit rule leads to slightly higher selection consistency.

Absolute Runtime (AR) On average it took 2636.756 model estimations when using sub-optimal heuristic information, before reaching either of the abort criteria. The longest run required 50528 total CFAs to be estimated (utilizing $\alpha = 1$, $\beta = 1$, $\rho = 0.95$, $K = 32$, and the *ib*

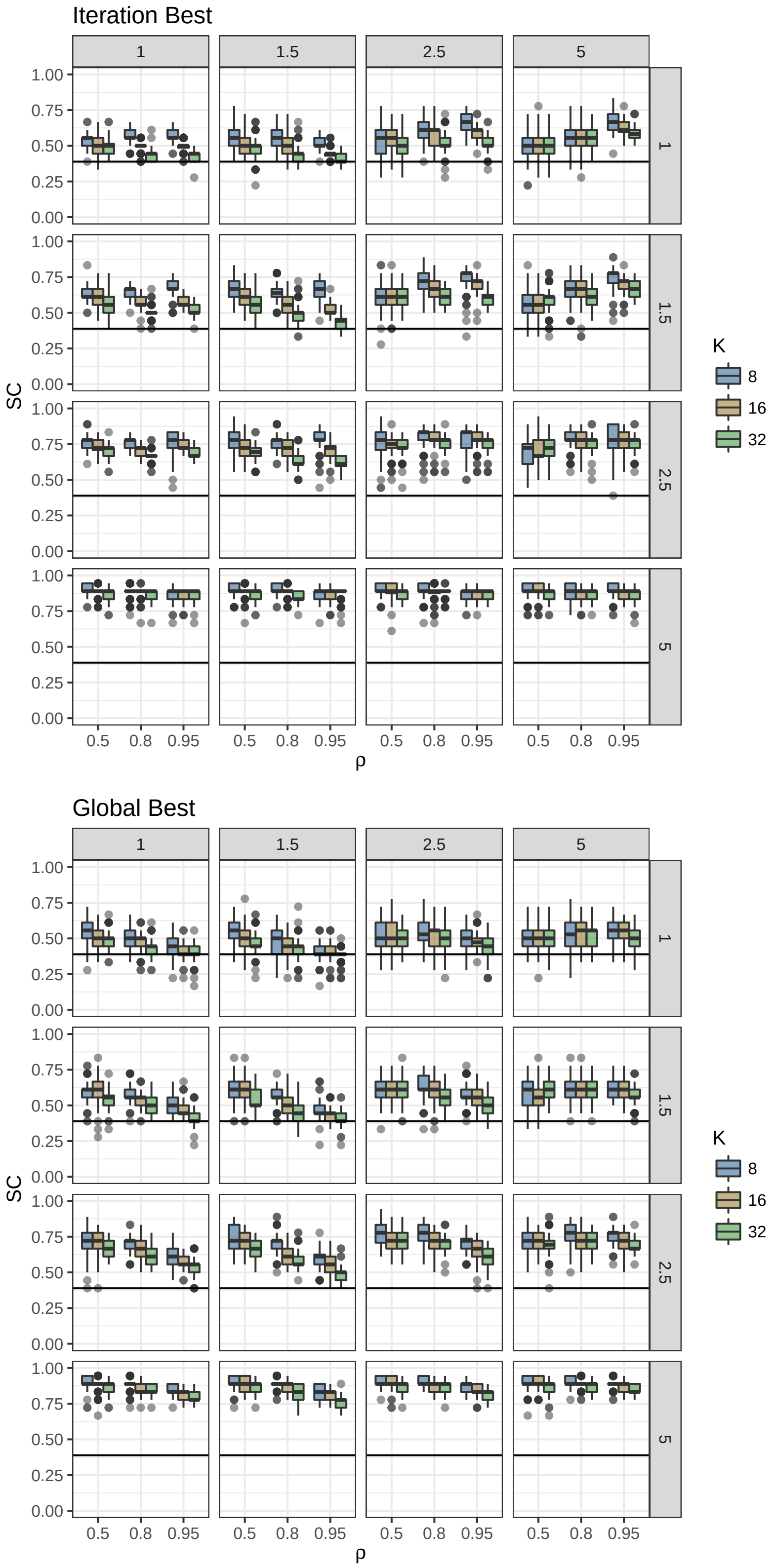


Figure 3.20: Boxplots of the selection consistency (SC) between the final solutions and the heuristically favored median solution. Panels represent different values for α in columns and β in rows. The solid line represents the SC of the optimal solution.

deposit rule). Though this took 195.878 minutes to finish during the evaluation study, it should be noted that this represents running roughly 0.000015% of all possible combinations.

Figure 3.21 shows the boxplots of the absolute runtime across all conditions. Conditions utilizing the iteration best deposit required more than twice the number of CFA estimations before reaching an abort criterion ($AR_{ib} = 3706.689$, $AR_{gb} = 1567.27$, Cohen's $d = 0.541$, 95% $CI = [0.517; 0.564]$). Despite this, all other factors, except for β, showed larger effects on AR when using ΔR^2 as an effect size estimate. Chief among them was the effect of the evaporation coefficient ρ ($\Delta R^2 = 0.423$, 95% $CI = [0.414; 0.432]$), for which larger values lead to much longer runtimes - with the average runtime for conditions with $\rho = .95$ being almost tenfold of those with $\rho = .5$. For the non-linearity coefficient α, larger values lead to shorter AR ($\Delta R^2 = 0.34$, 95% $CI = [0.331; 0.348]$), while more ants per colony lead to longer absolute runtimes ($\Delta R^2 = 0.277$, 95% $CI = [0.269; 0.285]$).

Overall, 925 (3.212%) replications required the minimum runtime determined via Equation (3.9), out of which 749 (80.973%) utilized either $\beta = 5$ or $\beta = 2.5$.

Relative Exploration (RE) Across all conditions the average $RE = 0.321$, indicating that less than a third of all generated solutions were unique and viable. Using ΔR^2 of the factors and all their respective interactions as the effect size measure, both non-linearity coefficients were the critical elements in determining RE values ($\Delta R^2_\alpha = 0.564$, 95% $CI = [0.555; 0.573]$, $\Delta R^2_\beta = 0.423$, 95% $CI = [0.415; 0.431]$). For both coefficients, higher values lead to considerably less exploration. The importance of α for RE is indicated by the fact that the 47 conditions with the highest values of RE are all conditions with $\alpha = 1$.

Higher values of the evaporation coefficient ρ generally lead to less exploration, though Figure 3.22 indicates that there may be no such trend in the best performing conditions (i.e. when $\alpha = 1$ and $\beta \in \{1, 1.5\}$) using the global-deposit rules. Despite the fact that, in terms of peak performance, the iteration-best outperformed the global-best deposit rule (i.e. the 10 conditions with the highest RE all utilize ib), on average the global-best deposit rule results in slightly more exploration ($RE_{ib} = 0.288$, $RE_{gb} = 0.354$, Cohen's $d = -0.306$, 95% $CI = [-0.329; -0.283]$).

The findings regarding the effects of K on relative exploration are mixed. Across all conditions the effect is negligible ($\Delta R^2 = 0.007$, 95% $CI = [0.004; 0.009]$). However, there was a small tendency for larger values of K to lead to less exploration, when using the global-best deposit rule ($RE_{K=8} = 0.376$, $RE_{K=16} = 0.353$, $RE_{K=32} = 0.334$), while the RE values were unsystematically varying for conditions with the iteration-best deposit rule.

The highest relative exploration achieved was $RE = 0.822$ in the condition with $\alpha = 1$, $\beta = 1$, $\rho = 0.5$, $K = 8$ and utilizing the iteration-best deposit rule. The six conditions with the largest RE values all used $\alpha = 1$, $\beta = 1$, $K \in \{8, 16\}$, and the iteration best deposit rule, though it should be noted that 29 conditions showed $RE > .7$ utilizing any number of ants per colony and any evaporation coefficient.

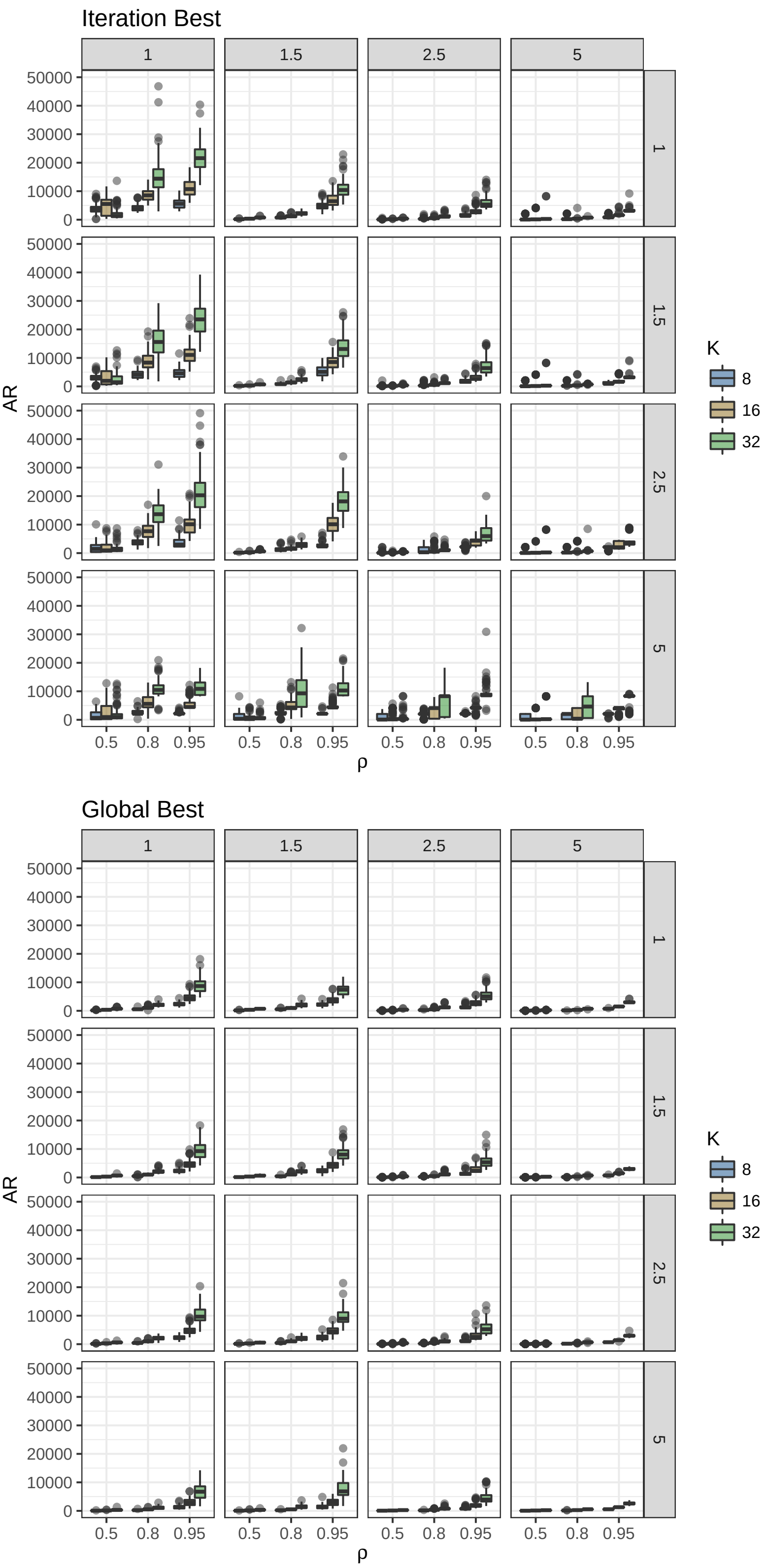


Figure 3.21: Boxplots of the absolute runtime (AR) for all conditions with sub-optimal heuristics. Panels represent different values for α in columns and β in rows.

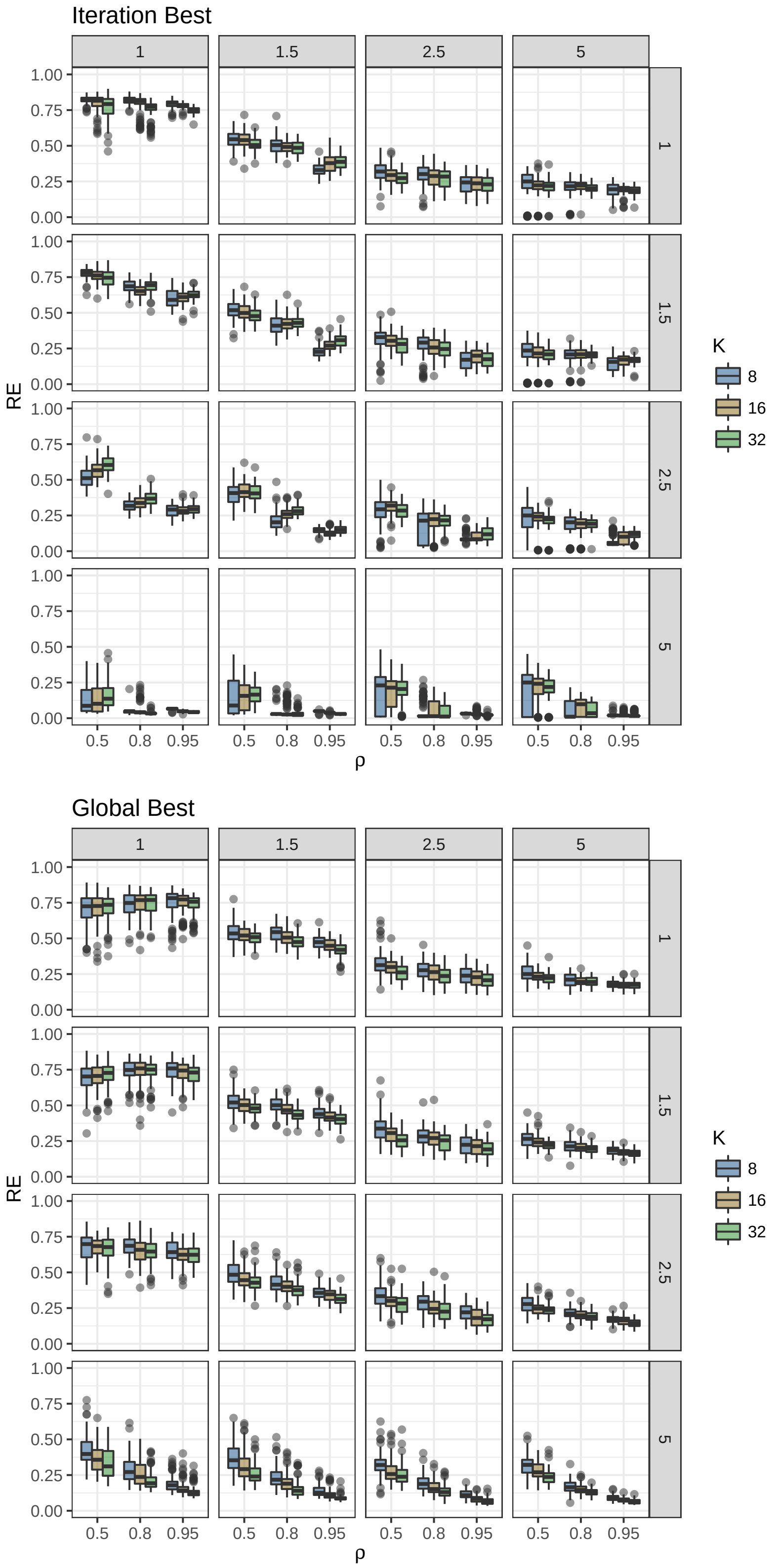


Figure 3.22: Boxplots of the relative exploration (RE) for all conditions with sub-optimal heuristics. Panels represent different values for α in columns and β in rows.

Table 3.14: Standard deviations of the current s^{gb} for different settings of α and β.

Deposit	α	$\beta = 1$	$\beta = 1.5$	$\beta = 2.5$	$\beta = 5$	M_{SD}
ib	1	0.042	0.043	0.038	0.038	0.040
	1.5	0.056	0.053	0.048	0.042	0.050
	2.5	0.062	0.054	0.040	0.044	0.050
	5	0.074	0.064	0.046	0.042	0.057
	M_{SD}	0.059	0.054	0.043	0.041	0.049
gb	1	0.065	0.062	0.057	0.049	0.058
	1.5	0.065	0.062	0.060	0.054	0.060
	2.5	0.067	0.063	0.058	0.052	0.060
	5	0.070	0.064	0.057	0.050	0.060
	M_{SD}	0.067	0.063	0.058	0.051	0.060

Relative Deviation over Time ($RD[t]$) Figures 3.23 through 3.26 show the $RD(t)$ of all replications that terminated normally. Be aware, that due to the drastic differences in absolute runtime between some of the conditions, the figures utilize different x-Axes and all are log-scaled.

Figures 3.23 and 3.24 depict the $RD(t)$ of the conditions utilizing the iteration-best deposit rule. Among these, higher values of β lead to a less steep trajectory of the relative deviation over time. This indicates current best solutions that are more similar in terms of the quality function and are generally closer to the solution favored in the heuristic information (as shown in the previous section). Larger values of α lead to less iterative updates of the currently best solution - and therefore to quicker convergence. This effect does not seem as pronounced in those conditions using the global-best deposit rule (depicted in Figures 3.25 and 3.26).

On average, 73.51 iterations were required until encountering a final solution, though this varied widely across conditions (range: [1;1323]). The final solution was replicated 71.298 times, on average, but again this number was vastly different across conditions with a range of [3;256]. It is also noteworthy, that increases in β lead to smaller variation in the quality of the current solutions across all conditions, while (at least when using the *ib* deposit rule) the opposite is true for larger values of α. Table 3.14 shows the standard deviations of the $RD(t)$ for different values of α and β. As indicated by the marginal means, the SD was higher across all conditions utilizing the *gb* deposit rule.

Relative Exploration over Time ($RE[t]$) Figures 3.27 and 3.28 depict the LOWESS lines for the $RE(t)$ across the 100 replications per condition. As was the case for the conditions without heuristic information, there was an initial burn-in phase for most conditions during which the $RE(t)$ increases due to the higher percentage of improper solutions during early iterations. However, all conditions with $\beta = 5$ do not show this burn-in, but instead exhibit constantly decreasing $RE(t)$. The same can be said about most conditions utilizing the iteration-best

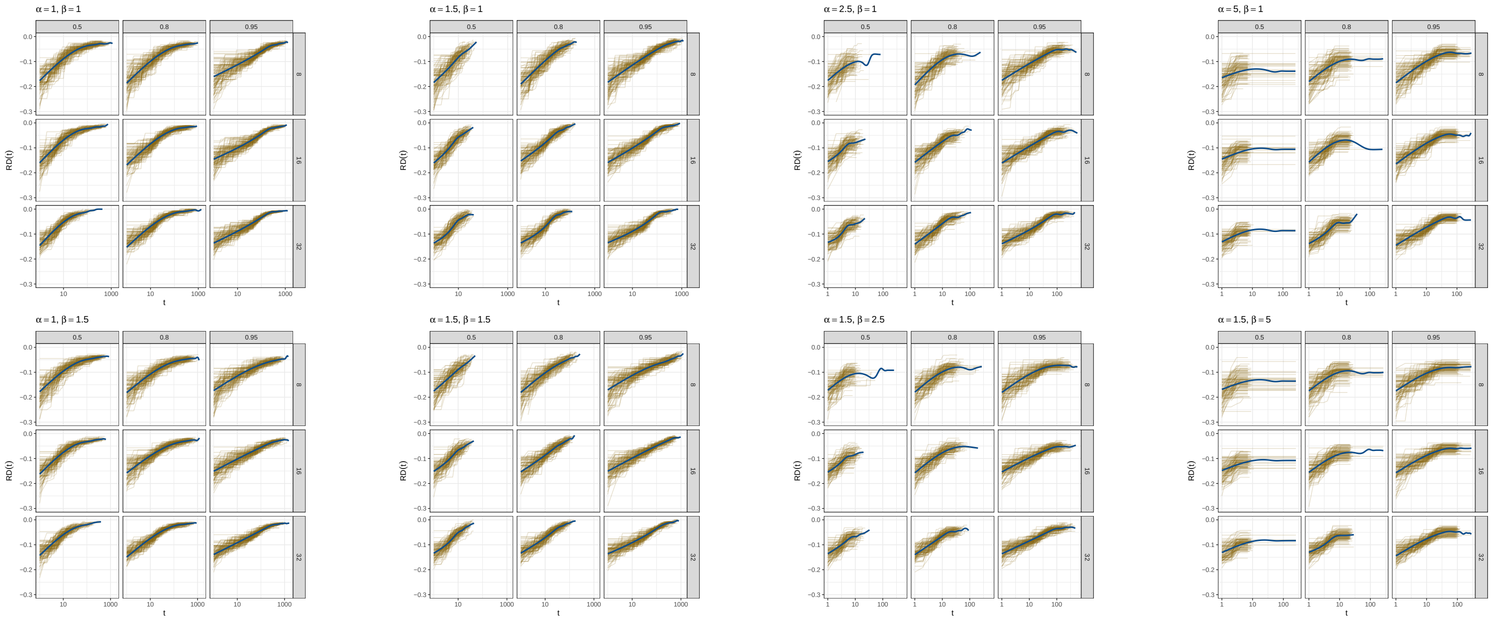


Figure 3.23: The relative deviation from $f(s^{opt})$ as a function of time ($RD[t]$) for the conditions using the iteration-best deposit rule and $\beta = 1$ or $\beta = 1.5$. Yellow lines represent the optimization history for each single replication, blue lines the LOESS-smoothed averages. The x-axis is log-scaled. The columns of each panel show different values of ρ, the rows different values of K.

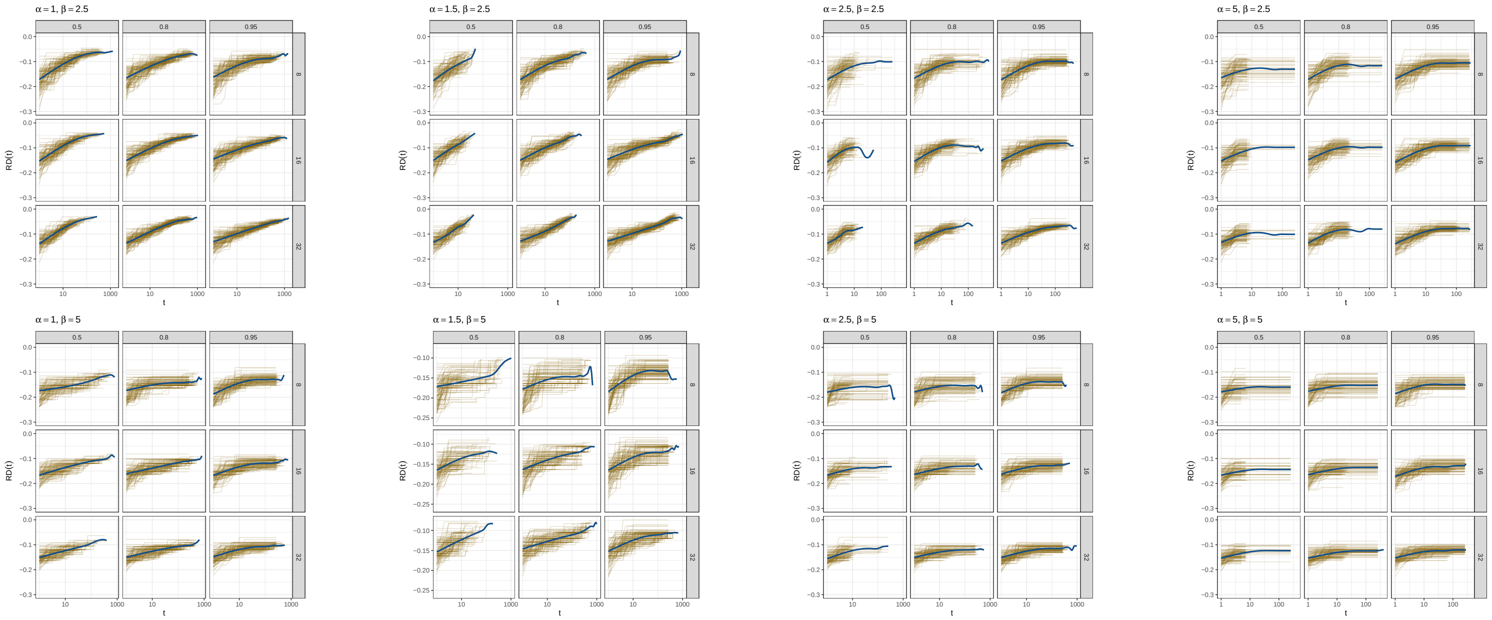


Figure 3.24: The relative deviation from $f(s^{opt})$ as a function of time $(RD[t])$ for the conditions using the iteration-best deposit rule and $\alpha = 2.5$ or $\alpha = 5$. Yellow lines represent the optimization history for each single replication, blue lines the LOESS-smoothed averages. The x-axis is log-scaled. The columns of each panel show different values of ρ, the rows different values of K.

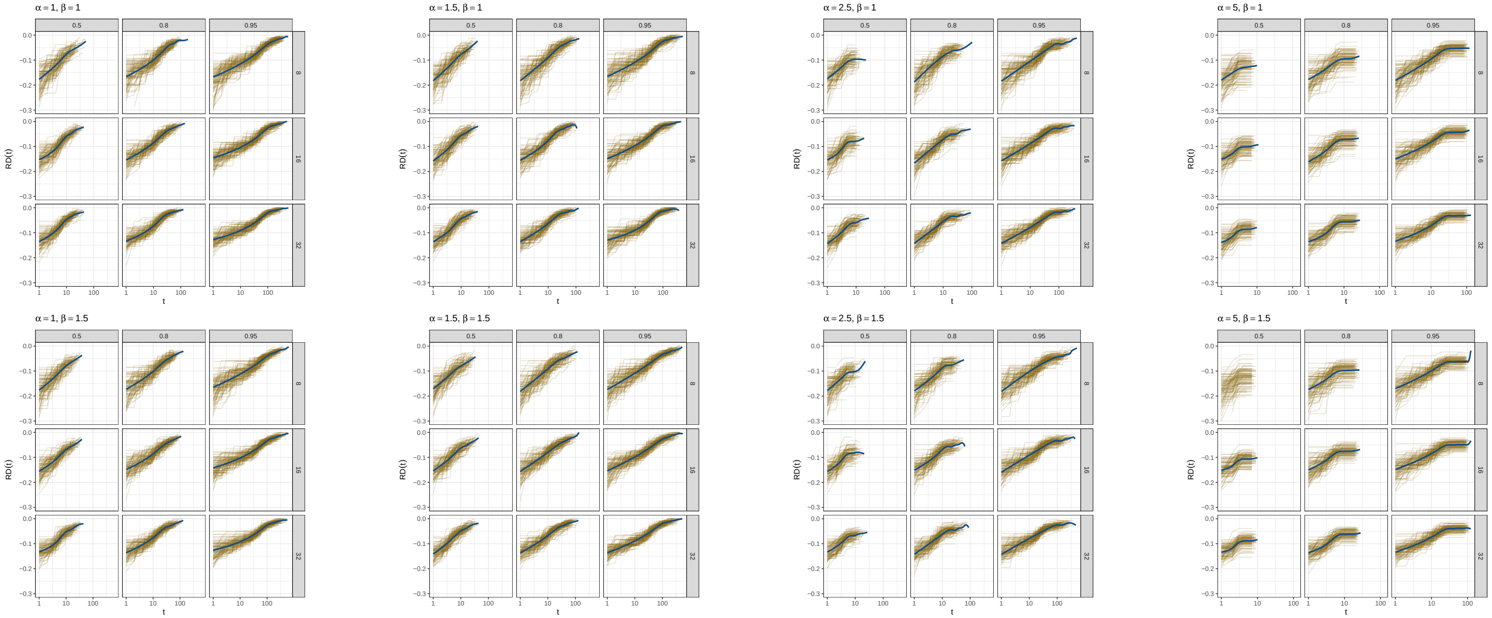


Figure 3.25: The relative deviation from $f(s^{opt})$ as a function of time ($RD[t]$) for the conditions using the global-best deposit rule and $\beta = 1$ or $\beta = 1.5$. Yellow lines represent the optimization history for each single replication, blue lines the LOESS-smoothed averages. The x-axis is log-scaled. The columns of each panel show different values of ρ, the rows different values of K.

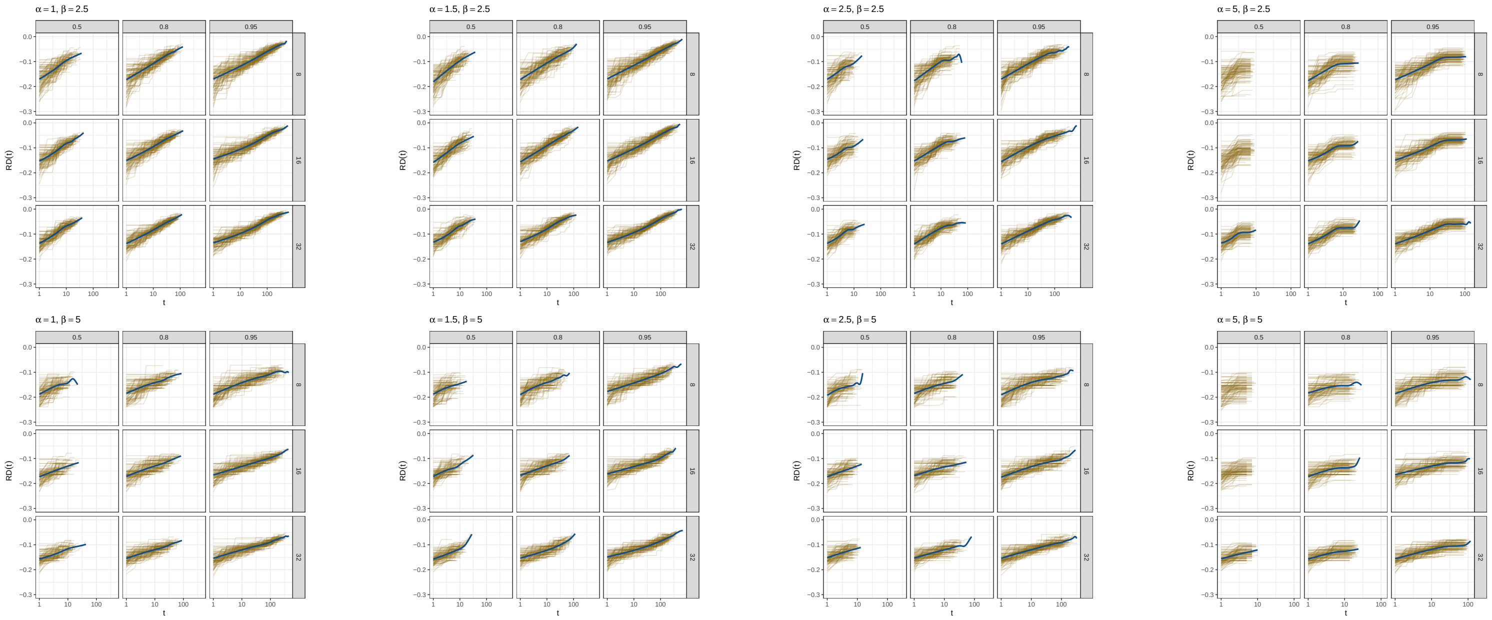


Figure 3.26: The relative deviation from $f(s^{opt})$ as a function of time ($RD[t]$) for the conditions using the global-best deposit rule and $\beta = 2.5$ or $\beta = 5$. Yellow lines represent the optimization history for each single replication, blue lines the LOESS-smoothed averages. The x-axis is log-scaled. The columns of each panel show different values of ρ, the rows different values of K.

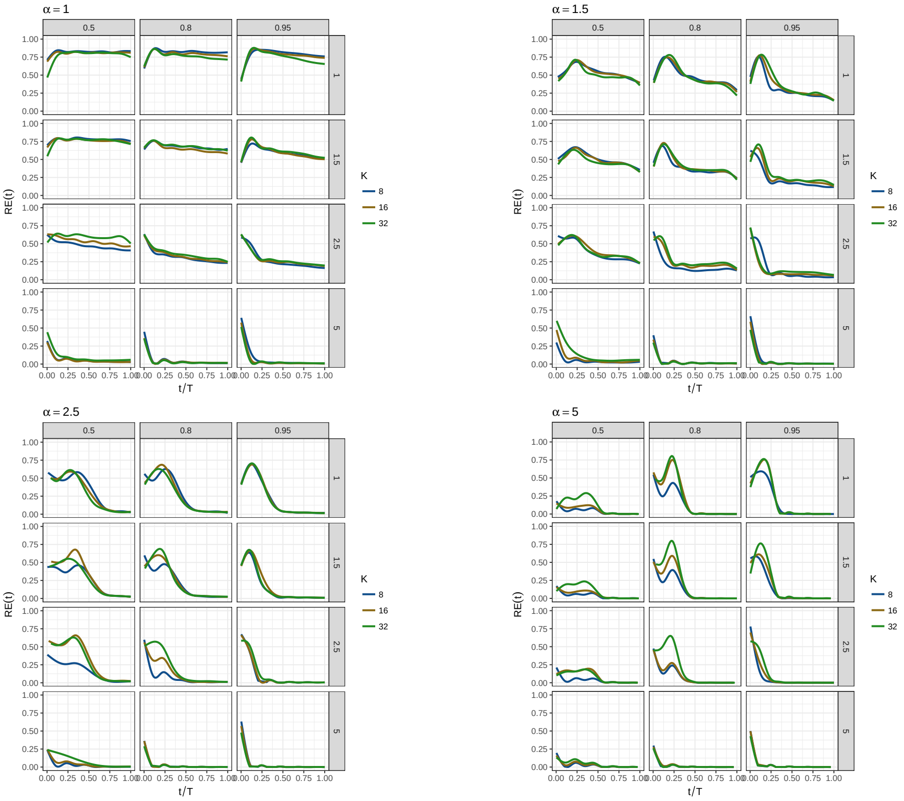


Figure 3.27: The LOWESS lines of the relative exploration over time for conditions with iteration-best deposit rule. Different values of ρ are in columns, different values of β in rows, and different values of α are in different panels.

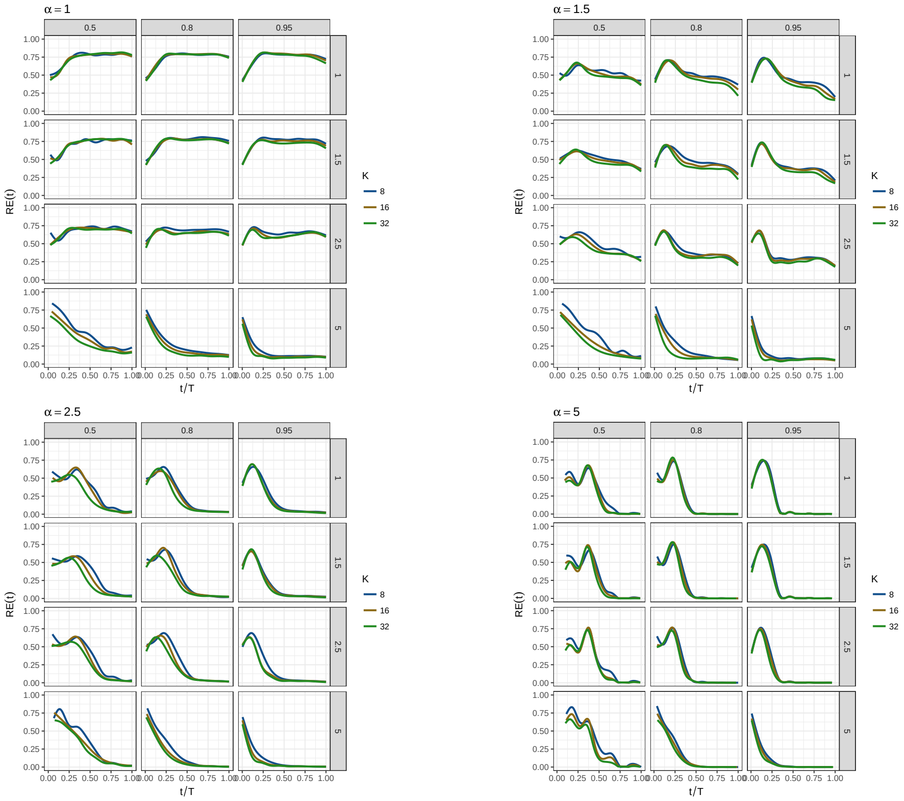


Figure 3.28: The LOWESS lines of the relative exploration over time for conditions with global-best deposit rule. Different values of ρ are in columns, different values of β in rows, and different values of α are in different panels.

deposit rule and $\beta = 2.5$, though at least four conditions in Figure 3.27 with this combination of parameters manifests some burn-in iterations.

Conditions with the *ib* deposit rule and $\beta = 5$ did not display much exploration after roughly 20% of iterations. In conditions utilizing the global-best deposit rule, this lack of exploration was also noticeable though far less pronounced. Irrespective of the deposit rule, this effect was amplified by larger values of ρ. When $\alpha = 1$ relative exploration is somewhat constant over time for conditions with $\beta \in \{1, 1.5\}$ and the iteration-best deposit rule and for those with $\beta \neq 5$ and the global-best deposit rule. Overall, though RE was higher on average for conditions with *ib* deposit as shown above, $RE(t)$ was more stable for conditions which use *gb* deposit.

Tables 3.15 and 3.16 show the average number required to reach the 90th percentile of $RE(t)$. Results for conditions with $\beta = 5$ again indicate steadily declining $RE(t)$ with the 90th percentile being reached after an average of 1.044 iterations when employing the *ib* deposit rule and after 1.101 iterations for conditions with the *gb* deposit rule. Additionally, the 90th percentile of $RE(t)$ for conditions with $\beta = 5$ and the *ib* deposit rule never exceeded 0.282.

Overall, conditions with $\alpha = 1$ and smaller values of K surmounted the burn-in phase quickly and approached very high levels of $RE(t)$ - going as far as constructing 100% unique and viable solutions, when using the *ib*, and up to 96.2% when using the *gb* deposit rule. As pointed out with regards to the figures, peak performance of the conditions utilizing the iteration-best deposit rule was better (Cohen's $d = 0.44$, 95% $CI = [0.463; 0.417]$) while there does not seem to be a meaningful difference with regards to the time required to overcome the burn-in (Cohen's $d = 0.069$, 95% $CI = [0.092; 0.045]$). However, specifically with regards to those conditions that attained very good $RE(t)$ 90th percentiles, i.e. conditions with $\alpha = 1$ and $\beta \in \{1, 1.5\}$, the global-best deposit rule seemed to attain comparable performance in less iterations.

Summary Across conditions with sub-optimal heuristics, higher values of β lead to results more in line with the solution provided by heuristic information, at the cost of the overall solution quality. With regards to α, higher values decreased the overall quality of the final solution without making it more consistent with the heuristic information. However, higher values of α lead to shorter runtimes - at the expense of less exploration. The non-linearity coefficient β had much less influence on runtimes, but decreased overall exploration nonetheless. With regards to the evaporation coefficient ρ results are somewhat unclear. While lower values of ρ lead to lower solution quality in many cases, they also lead to higher average exploration in shorter runtimes. The selection of a deposit rule can have dramatic effects on runtimes and solution quality, depending on the settings of α and β. Generally, the *gb* deposit rule seems more resistant to heuristic information - leading to higher solution quality and lower selection consistency when using the same settings for the non-linearity coefficients.

Table 3.15: Average number of iterations needed for the burn-in phase (T_{bi}) per condition with the average 90th percentile of RE in parentheses.

β	K	ρ	$\alpha = 1$		$\alpha = 1.5$		$\alpha = 2.5$		$\alpha = 5$	
1	8	0.5	10.684	(1.000)	4.182	(0.799)	3.175	(0.668)	1.907	(0.074)
		0.8	19.410	(1.000)	7.265	(0.815)	5.566	(0.699)	2.860	(0.384)
		0.95	47.653	(0.996)	5.040	(0.680)	9.828	(0.739)	5.677	(0.611)
	16	0.5	11.840	(0.953)	4.580	(0.769)	3.110	(0.667)	2.400	(0.120)
		0.8	19.070	(0.928)	9.120	(0.789)	6.140	(0.758)	3.740	(0.638)
		0.95	55.320	(0.934)	18.370	(0.714)	14.220	(0.714)	8.780	(0.716)
	32	0.5	10.670	(0.907)	4.760	(0.712)	3.350	(0.638)	2.540	(0.266)
		0.8	17.810	(0.892)	9.690	(0.784)	6.400	(0.711)	3.740	(0.676)
		0.95	56.480	(0.886)	26.520	(0.751)	17.300	(0.717)	10.060	(0.715)
1.5	8	0.5	11.000	(0.998)	3.837	(0.788)	2.768	(0.558)	1.980	(0.070)
		0.8	11.070	(0.900)	4.449	(0.700)	4.340	(0.563)	2.571	(0.350)
		0.95	18.450	(0.827)	2.414	(0.509)	4.788	(0.569)	3.630	(0.461)
	16	0.5	9.900	(0.930)	4.050	(0.725)	3.100	(0.686)	2.350	(0.109)
		0.8	12.230	(0.844)	6.170	(0.701)	5.190	(0.670)	3.380	(0.509)
		0.95	27.060	(0.800)	4.390	(0.558)	8.650	(0.645)	5.940	(0.556)
	32	0.5	10.740	(0.892)	4.150	(0.684)	3.240	(0.632)	2.610	(0.231)
		0.8	13.880	(0.834)	7.370	(0.691)	5.530	(0.710)	3.690	(0.678)
		0.95	36.440	(0.780)	11.410	(0.604)	10.710	(0.608)	8.130	(0.674)
2.5	8	0.5	3.398	(0.753)	2.449	(0.679)	2.560	(0.397)	1.909	(0.054)
		0.8	1.680	(0.558)	1.444	(0.418)	2.390	(0.200)	2.172	(0.204)
		0.95	1.770	(0.549)	1.459	(0.471)	1.200	(0.367)	1.242	(0.079)
	16	0.5	4.510	(0.760)	2.990	(0.679)	3.020	(0.704)	2.280	(0.195)
		0.8	2.230	(0.530)	1.950	(0.504)	3.600	(0.382)	2.520	(0.239)
		0.95	1.890	(0.484)	1.130	(0.273)	1.770	(0.385)	1.990	(0.207)
	32	0.5	6.290	(0.760)	3.350	(0.655)	2.930	(0.669)	2.360	(0.179)
		0.8	2.750	(0.529)	2.740	(0.526)	3.900	(0.593)	3.130	(0.579)
		0.95	2.150	(0.461)	1.030	(0.293)	3.350	(0.448)	2.690	(0.372)
5	8	0.5	1.000	(0.149)	1.090	(0.149)	1.170	(0.038)	1.230	(0.035)
		0.8	1.000	(0.125)	1.000	(0.111)	1.000	(0.014)	1.010	(0.018)
		0.95	1.000	(0.226)	1.000	(0.130)	1.000	(0.030)	1.000	(0.003)
	16	0.5	1.040	(0.137)	1.080	(0.158)	1.170	(0.088)	1.250	(0.053)
		0.8	1.010	(0.095)	1.000	(0.072)	1.000	(0.025)	1.010	(0.043)
		0.95	1.000	(0.119)	1.000	(0.079)	1.000	(0.054)	1.000	(0.016)
	32	0.5	1.030	(0.184)	1.040	(0.282)	1.220	(0.196)	1.210	(0.086)
		0.8	1.000	(0.067)	1.010	(0.044)	1.010	(0.018)	1.020	(0.029)
		0.95	1.000	(0.086)	1.000	(0.047)	1.000	(0.026)	1.000	(0.009)

Table 3.16: Average number of iterations needed for the burn-in phase (T_{bi}) per condition with the average 90th percentile of RE in parentheses.

β	K	ρ	$\alpha = 1$		$\alpha = 1.5$		$\alpha = 2.5$		$\alpha = 5$	
1	8	0.5	6.414	(0.935)	4.280	(0.775)	2.800	(0.683)	2.100	(0.623)
		0.8	13.650	(0.947)	6.469	(0.808)	4.229	(0.712)	2.375	(0.674)
		0.95	31.430	(0.954)	10.847	(0.789)	6.545	(0.701)	4.384	(0.701)
	16	0.5	8.250	(0.902)	4.440	(0.730)	3.200	(0.675)	2.500	(0.580)
		0.8	15.410	(0.909)	7.730	(0.751)	4.730	(0.672)	3.210	(0.646)
		0.95	46.430	(0.913)	17.860	(0.760)	10.910	(0.689)	5.900	(0.694)
	32	0.5	9.890	(0.881)	4.340	(0.694)	3.010	(0.603)	2.770	(0.547)
		0.8	17.150	(0.881)	7.770	(0.720)	5.170	(0.657)	3.660	(0.654)
		0.95	50.830	(0.882)	22.200	(0.737)	12.960	(0.669)	7.670	(0.681)
1.5	8	0.5	6.612	(0.922)	3.700	(0.774)	2.830	(0.686)	2.030	(0.641)
		0.8	13.979	(0.960)	6.270	(0.797)	4.061	(0.741)	2.960	(0.680)
		0.95	30.860	(0.962)	9.111	(0.759)	6.510	(0.675)	4.220	(0.719)
	16	0.5	7.370	(0.884)	4.040	(0.712)	3.020	(0.661)	2.570	(0.589)
		0.8	15.220	(0.908)	6.540	(0.727)	4.940	(0.716)	3.110	(0.683)
		0.95	40.890	(0.893)	12.430	(0.704)	8.240	(0.630)	5.690	(0.677)
	32	0.5	9.060	(0.869)	4.090	(0.665)	2.950	(0.610)	2.710	(0.550)
		0.8	16.900	(0.877)	6.590	(0.678)	4.830	(0.649)	3.330	(0.651)
		0.95	45.240	(0.852)	18.030	(0.710)	10.340	(0.638)	6.500	(0.647)
2.5	8	0.5	5.210	(0.899)	2.930	(0.757)	2.770	(0.743)	2.061	(0.671)
		0.8	8.620	(0.904)	3.929	(0.738)	3.606	(0.732)	2.290	(0.704)
		0.95	15.747	(0.885)	3.890	(0.677)	4.660	(0.676)	3.030	(0.690)
	16	0.5	5.880	(0.847)	3.330	(0.693)	2.850	(0.677)	2.420	(0.608)
		0.8	9.620	(0.828)	4.430	(0.679)	4.180	(0.702)	2.890	(0.679)
		0.95	19.530	(0.813)	5.570	(0.634)	4.760	(0.568)	3.980	(0.653)
	32	0.5	6.950	(0.819)	3.320	(0.661)	2.830	(0.647)	2.660	(0.582)
		0.8	11.910	(0.803)	4.370	(0.643)	3.710	(0.641)	2.890	(0.640)
		0.95	24.530	(0.773)	5.370	(0.585)	5.750	(0.562)	4.170	(0.622)
5	8	0.5	1.270	(0.715)	1.220	(0.692)	1.300	(0.699)	1.210	(0.729)
		0.8	1.160	(0.550)	1.080	(0.527)	1.050	(0.562)	1.060	(0.565)
		0.95	1.030	(0.411)	1.010	(0.342)	1.010	(0.362)	1.000	(0.339)
	16	0.5	1.300	(0.593)	1.170	(0.573)	1.180	(0.615)	1.350	(0.654)
		0.8	1.080	(0.472)	1.060	(0.432)	1.030	(0.462)	1.050	(0.499)
		0.95	1.000	(0.314)	1.000	(0.261)	1.010	(0.243)	1.000	(0.283)
	32	0.5	1.240	(0.535)	1.230	(0.535)	1.210	(0.575)	1.250	(0.604)
		0.8	1.020	(0.402)	1.000	(0.310)	1.020	(0.407)	1.040	(0.467)
		0.95	1.000	(0.234)	1.000	(0.181)	1.000	(0.180)	1.000	(0.210)

3.8 Parameter Schedules

The results shown in Section 3.7 are used in this section to elicit parameter schedules which appear promising in terms of performance and should therefore be investigated in more detail. As discussed in Section 3.1.8, parameter schedules can often enhance the speed of ACO algorithms immensely, because they allow for more specific behavior during different phases of the search procedure. Section 3.7 revealed that early stages of the search tend to be characterized by a burn-in phase which is needed to separate inadmissible solutions from those contained within the set $\mathcal{S}^*$. During this stage it seems advisable to use parameter settings which reach high levels of $RE(t)$ quickly. In contrast, settings which do not converge too fast and are instead characterized by higher $RD(t)$ are desirable during the broad search for good solutions in $\mathcal{S}^*$, because many solutions may be viable, albeit not close to optimal. After close to optimal solutions are identified, another shift in parameters may be advisable to ensure a more localized exploration and quick evaporation of pheromones on items leading to worse solutions.

Thus, this section will aim to identify promising combinations of parameters, as derived from results with constant parameter settings, and schedule them accordingly. In contrast to the evaluation of constant parameters settings, this section will not include settings with optimal heuristic information. In the previous evaluation these conditions were included to investigate whether the approach presented in this thesis can recover optimal solutions sufficiently often when given the optimal conditions to run in. Because it is the aim of this part of the parameter evaluation to culminate in explicit recommendations for applications of this approach, it is not necessary to consider situations in which optimal solutions are known. Instead the focus will be on situations in which either no or sub-optimal heuristics are provided.

A general comparison, included in both evaluations irrespective of whether heuristic information is provided, will be performed on the type of parameter scheduling. As discussed in Section 3.1.8 different types of parameter scheduling can lead to different search behavior. The first type utilized in this evaluation is a rigid parameter schedule (rps) which provides changes in parameters after a set number of iterations has passed. The second type is a somewhat more flexible parameter scheduling (fps) scheme, which resets its current phase every time a new s^{gb} is found.

For example, if one were to set $\alpha = 2.5$ for the first 10 iterations, to $\alpha = 1.5$ for the subsequent 10, and then $\alpha = 1$ for all following iterations, the rigid schedule would result in exactly 10 iterations being performed with $\alpha_1 = 2.5$, then exactly ten iterations with α_2 and all subsequent iterations with $\alpha_2 = 1$. In contrast, the flexible parameter schedule uses the number of iterations since the last s^{gb} was found (denoted t^{gb}) to reinitialize the current phase. This way, if no s^{gb} were found in the first 10 iterations, $\alpha_2 = 1.5$ would be initialized as the second phase. If a new s^{gb} were then found at $t = 14$, $t^{gb} = 1$ would be set and the number of iterations to reach the switch to α_3 would be ten ($t_{\alpha_3} - t_{\alpha_2} = 20 - 10 = 10$) from this point, assuming that no new s^{gb} is found during these ten iterations.

While previous studies showed that schedules akin to the rigid parameter schedule tend to perform better than more flexible approaches (see Section 3.1.8 for an overview), they are harder to define for each application. Setting schedules requires knowledge about problem size and its influence on the behavior of the search. Especially in item selection this is extremely speculative before any search has been initialized, because the number of viable solutions has a major impact on the performance of the search. With this insecurity, it may be easier to define fps than rps in an application. Because of this expected benefit of the fps, their comparison is of relevance to the recommendations for application of the `stuart` approach.

3.8.1 Scheduled Parameters with No Heuristic Information

Section 3.7.1 showed some clear general trends in the effects of certain parameters on solution quality and search speed in the construction of short-scales: (a) values of $\alpha > 1$ are not recommendable, (b) the iteration-best deposit rule generally outperforms the global-best deposit rule in terms of peak performance, and (c) lower values of ρ and K lead to a quicker search.

As pointed out with regards to α, values other than 1 generally lead to very fast convergence at the expense of reduced exploration and lower quality of the final solution. While the burn-in phase regarding $RE(t)$ is generally shorter for conditions with $\alpha > 1$, the peak $RE(t)$ is unable to reach the values found in conditions with $\alpha = 1$. Results regarding the $RD(t)$ indicate no advantage of using settings other than $\alpha = 1$. Thus, in this evaluation only conditions with $\alpha = 1$ are considered.

In terms of solution quality, utilizing the *ib* deposit rule generally proved better than using its *gb* counterpart. Additionally, the iteration-best deposit results in higher and more consistent RE than the global-best when using $\alpha = 1$. The only tangible benefit of using the *gb* deposit rule in situations without heuristic information is the decrease in AR. However, because this decrease in runtime comes at a cost to solution quality and exploration, the *ib* deposit rule is considered superior when used in the absence of heuristic information. Therefore, this evaluation will investigate only conditions utilizing the *ib* deposit rule.

Results concerning the number of ants per colony (K) and the evaporation coefficient ρ revealed better peak performance in conditions in which both are high. However, results regarding the timed performance criteria $RD(t)$ and $RE(t)$ revealed conditions with low vales in K and ρ to surmount the burn-in phase quickly and then result in very high relative exploration. Therefore, beginning the algorithm with low settings of K and ρ seems a promising approach to reach a phase of exploration in $\mathcal{S}^*$ quickly and reliably. K and ρ are scheduled to increase incrementally from 8 to 16 to 32 and from .5 to .8 to .95, respectively, in this evaluation. Because the final solutions show better values for RD when K and ρ are large, these parameter settings should be used during the middle exploration phase of the search.

Table 3.17: Parameter schedules for the evaluation with no heuristic information.

Parameter	Time	Condition	Values	Schedule	
				rps	*fps*
ρ	start	early	(.5,.8,.95)	(0,15,30)	(0,15,30)
		late	(.5,.8,.95)	(0,30,60)	(0,30,60)
		none	.95	-	-
	end	early	(.8,.5)	(150,200)	(75,150)
		late	(.8,.95)	(300,400)	(100,200)
		none	-	-	-
K	start	early	(8,16,32)	(0,15,30)	(0,15,30)
		late	(8,16,32)	(0,30,60)	(0,30,60)
		none	32	-	-
	end	early	(16,8)	(150,200)	(75,150)
		late	(16,8)	(300,400)	(100,200)
		none	-	-	-

As shown in Table 3.8 conditions utilizing the *ib* deposit rule and $\alpha = 1$ tend to require an average of about 12 to 20 iterations to surmount the burn-in phase. Thus, in a conservative condition, the two early parameter switches are scheduled for 15 and 30 iterations after initialization. In a more liberal approach these switches are scheduled for $t = 30$ and $t = 60$. In addition to these schedules, a constant condition is added which utilizes the largest values, i.e. $K = 32$ and $\rho = .95$, respectively. The schedules for the number of ants and the evaporation coefficient are varied independently, resulting in a fully crossed design. The same settings are used, irrespective of the scheduling type.

As pointed out with regards to the $RD(t)$, conditions with $\alpha = 1$ and $\rho = .95$ did not reach the abort criterion defined by the pheromone limits, but instead terminated only after the maximum number of iterations was reached. This indicates an unnecessarily long runtime due to a large number of iterations with no benefit to s^{gb} during the late stages of the search. To investigate whether convergence can be sped up, this evaluation includes conditions in which the number of ants and the evaporation coefficient are reduced during the later stages of the algorithm. These reductions are the mirrored opposite of the increases during the early stages (i.e. in sequence $k = [32, 16, 8]$ and $\rho = [.95, .8, .5]$). For conditions with the *rps* these will occur at $t = 300$ and $t = 400$ in one condition and at $t = 150$ and $t = 200$ in another. These values are chosen because, in conditions with $\alpha = 1$, the final s^{gb} was found after an average of roughly 270 iterations. To allow an adequate amount of time for near optimal solutions to be found, the conservative condition exceeds this average slightly before reducing ρ and K to facilitate convergence. The more liberal condition is defined under the assumption that the conditions with early shifts in the parameters will find near optimal solutions quicker than did conditions with constant parameter settings. Again, these two levels are complemented by a condition in which no scheduling is done during this phase of the search. In cases with the fps, the number of iterations is set to be $t = (100, 200)$ in a conservative, and $t = (75, 150)$ in a more liberal condition. These values are chosen so that they allow for an adequate number of replications of the final s^{gb} before enforcing convergence.

In a fully crossed design this results in 162 conditions. Table 3.17 provides an overview of the conditions used in this evaluation.

3.8.2 Results with No Heuristic Information

Errors Over all conditions including no heuristic information and scheduled parameter settings, 129 of 16200 replications (0.796%) resulted in errors. All of these errors were due to no viable solution being found in the first iteration, which occurred almost equally often in conditions with early or late beginning schedules for the number of ants (57 vs. 72). Both of these conditions began with $K = 8$.

Success Rate (SR) On average, the SR of conditions with no heuristic information and the scheduled parameters settings was 0.13. Table 3.18 provides more details on the condition specific SR. Over all conditions, differences in the ending schedule of the number of ants K had the greatest impact on SR ($\Delta R^2 = 0.034$, 95% $CI = [0.027; 0.04]$), while no other influence showed $R^2 > 0.02$. Most notably, conditions using an early reduction of ants had lower SR than the other two conditions at $SR = 0.069$. Across all settings for K and ρ there were only negligible differences in SR between the scheduling types (Cohen's $d = -0.124$, 95% $CI = [-0.154; -0.093]$).

Relative Deviation (RD) On average, the conditions under investigation in this section achieved an $RD = -0.005$, meaning that the *average* performance of the final solutions constructed here was just 0.527% worse than the optimum. Figure 3.29 shows the RD of the conditions in more detail.

Regarding the scheduling of K, Figure 3.29 shows that the end schedules performed worse, the quicker the number of ants was reduced ($RD_{\text{early}} = -6.29 \times 10^{-3}$, $RD_{\text{late}} = -4.93 \times 10^{-3}$, $RD_{\text{none}} = -4.57 \times 10^{-3}$). This constitutes the largest determinant of RD in this evaluation ($\Delta R^2 = 0.03$, 95% $CI = [0.02; 0.04]$). The scheduling of K at the beginning of the algorithm had the opposite effect, with RD being best in situations in which many early iterations used a smaller number of ants ($RD_{\text{early}} = -4.85 \times 10^{-3}$, $RD_{\text{late}} = -4.72 \times 10^{-3}$, $RD_{\text{none}} = -6.21 \times 10^{-3}$). Of the ten conditions with the best RD values, nine used no schedule on K at the end of the search.

The schedules implemented for ρ had little effect on the average quality of solutions, they did, however, impact the variability of solution quality. In particular, the starting schedules for ρ, depicted in the rows of panels in Figure 3.29, had an impact on the variance of $f(s^{gb})$ in such a way that including lower values of ρ in the beginning of the search increased variability

Table 3.18: Condition-specific SR for cases with scheduled parameters and no heuristic information.

			K start	early			late			none		
Type	ρ start	ρ end	K end	early	late	none	early	late	none	early	late	none
rps	early	early		0.020	0.144	0.150	0.031	0.141	0.110	0.060	0.150	0.130
		late		0.020	0.141	0.220	0.030	0.155	0.101	0.010	0.200	0.220
		none		0.051	0.112	0.182	0.041	0.101	0.194	0.090	0.130	0.220
	late	early		0.062	0.111	0.081	0.020	0.130	0.090	0.030	0.100	0.130
		late		0.051	0.121	0.100	0.060	0.091	0.306	0.070	0.110	0.220
		none		0.051	0.141	0.190	0.030	0.070	0.286	0.100	0.140	0.190
	none	early		0.030	0.160	0.153	0.000	0.172	0.143	0.020	0.120	0.050
		late		0.010	0.101	0.253	0.000	0.041	0.230	0.010	0.100	0.190
		none		0.010	0.071	0.202	0.011	0.081	0.202	0.000	0.050	0.210
fps	early	early		0.150	0.220	0.170	0.110	0.162	0.182	0.060	0.100	0.160
		late		0.190	0.182	0.110	0.101	0.186	0.227	0.090	0.140	0.110
		none		0.111	0.158	0.232	0.111	0.202	0.162	0.100	0.200	0.170
	late	early		0.040	0.182	0.155	0.124	0.170	0.155	0.090	0.080	0.060
		late		0.133	0.190	0.140	0.110	0.303	0.153	0.050	0.090	0.100
		none		0.101	0.133	0.212	0.090	0.224	0.263	0.070	0.130	0.150
	none	early		0.071	0.220	0.265	0.082	0.194	0.265	0.180	0.160	0.260
		late		0.090	0.200	0.190	0.071	0.200	0.152	0.170	0.220	0.170
		none		0.091	0.173	0.182	0.051	0.141	0.153	0.140	0.160	0.190

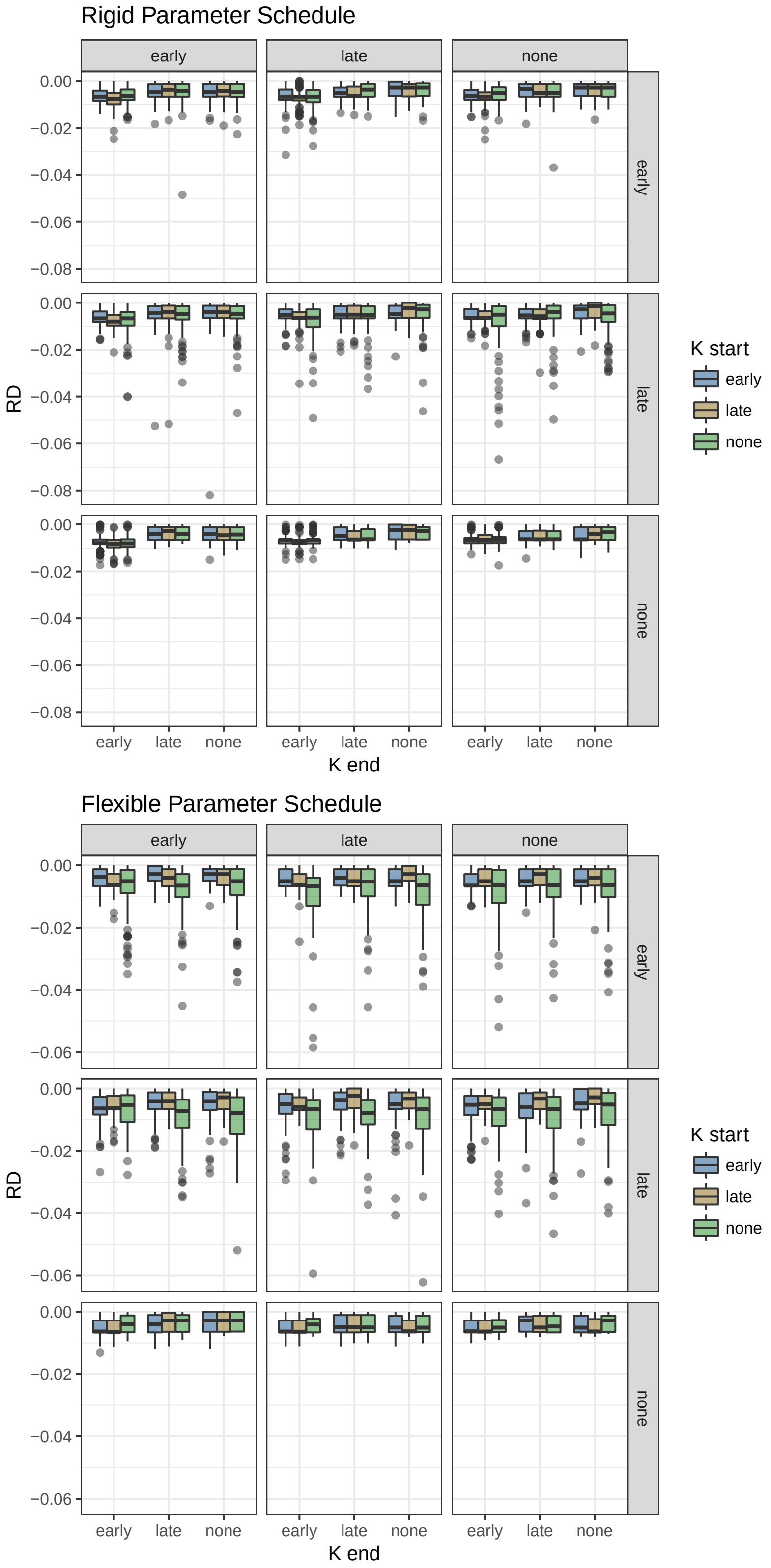


Figure 3.29: Boxplots of the relative deviation (RD) of the final solutions from the optimal solution. Panels of the figure are for different combinations of ρ schedules. The starting schedules are depicted in rows, the ending schedules in columns.

Table 3.19: Average and worst-case RD for the five best performing conditions in each respect. The ranks pertain to the total of all 182 conditions without heuristic information.

					Average		Worst-Case			
Schedule	K start	K end	ρ start	ρ end	RD	Rank	RD	Rank	SR	AR
rps	late	none	none	late	-0.003	1	-0.008	5	0.230	13093.440
rps	early	none	none	late	-0.003	2	-0.011	33	0.253	13292.606
rps	late	none	late	none	-0.003	3	-0.018	100	0.286	16478.531
fps	early	late	early	early	-0.003	4	-0.012	44	0.220	11801.440
fps	early	none	early	early	-0.003	5	-0.013	59	0.170	11402.400
fps	none	none	none	early	-0.003	7	-0.007	1	0.260	15778.240
rps	none	none	none	late	-0.003	13	-0.007	1	0.190	13844.800
fps	late	none	none	early	-0.004	15	-0.008	5	0.265	14185.306
fps	none	none	none	none	-0.004	19	-0.007	4	0.190	17087.040
fps	none	late	none	none	-0.004	38	-0.007	1	0.160	17545.280

of the final solutions substantially ($var(RD_{\text{early}}) = 2.49 \times 10^{-5}$, $var(RD_{\text{late}}) = 3.71 \times 10^{-5}$, $var(RD_{\text{none}}) = 8.89 \times 10^{-6}$). Inspecting the worst-case performance reveals that the 21 conditions with the best worst-case RD all utilized no evaporation schedule at the beginning of the search.

Because of this, the best performing conditions in terms of average and worst-case performance showed only little overlap. Table 3.19 shows the best performing conditions with respect to their average and worst-case RD. Of these conditions, only two used a schedule for the number of ants during the end of the search (in both cases combined with the fps schedule) and only three used a schedule for ρ at the start of the search.

Absolute Runtime (AR) On average, 10491.368 CFAs were run per replication. Using the ΔR^2 approach described in Section 3.5, all of the factors in this study contributed substantially to the absolute runtime, with all $\Delta R^2 > 0.1$.

The most important determinant of AR was the schedule used for K at the end of the search ($\Delta R^2 = 0.254$, 95% $CI = [0.239; 0.27]$). As can be expected, quicker switches to smaller K lead to shorter runtimes in general. Conversely, the K schedule at the beginning of the search lead to larger AR when more iterations used lower K ($AR_{\text{late}} = 10974.847$, $AR_{\text{early}} = 10783.519$,$AR_{\text{none}} = 9725.268$). These effects lead to the third largest effect with a $\Delta R^2 = 0.14$ (95% $CI = [0.124; 0.157]$).

The second most important factor for AR was the schedule type ($\Delta R^2 = 0.152$, 95% $CI = [0.136; 0.168]$), with the flexible parameter schedule being much slower than the rigid schedule on average ($AR_{fps} = 11414.634$, $AR_{rps} = 9567.527$). Overall, the scheduling of ρ had the least influence on AR, with starting schedules accomplishing a $\Delta R^2 = 0.123$ (95% $CI = [0.106; 0.139]$) and ending schedules a $\Delta R^2 = 0.119$ (95% $CI = [0.103; 0.136]$). However, Figure 3.31 shows the condition specific boxplots of AR, depicting one of the results for early ρ schedules. As was the case for RD, utilizing schedules for ρ at the beginning of the search lead to inconsistent, highly variable results. Especially cases utilizing a scheduled version of ρ at the beginning of the search, no schedule for K, and the fps lead to some cases converging to suboptimal solutions quickly. Across all conditions, the fps had much higher variability in the results ($sd[AR_{fps}] = 6194.145$, $sd[AR_{rps}] = 3878.967$).

Relative Exploration (RE) Across all conditions without heuristics the average $RE = 0.817$, indicating that just over 80% of all constructed solutions were unique and viable. As was the case in the previous sections, utilizing a schedule for ρ during the beginning of the search increased the variability of the RE across replications. This did not, however, have a noteworthy impact on the average RE ($\Delta R^2 = 0.015$, 95% $CI = [0.004; 0.026]$).

In general, none of the manipulated variables had a large impact on the RE and the condition specific averages were in a small range between 0.764 and 0.864. The scheduling of K resulted in the largest ΔR^2. At the beginning of the algorithm not utilizing a schedule lead to lower RE than both of the scheduled conditions ($RE_{\text{none}} = 0.803$, $RE_{\text{early}} = 0.821$, $RE_{\text{late}} = 0.828$). Overall this resulted in a $\Delta R^2 = 0.045$ (95% $CI = [0.035; 0.054]$). Similarly, for ant schedules at the end of the search, reducing the number of ants quickly lead to higher relative exploration ($RE_{\text{none}} = 0.809$, $RE_{\text{early}} = 0.829$, $RE_{\text{late}} = 0.814$) with an effect size of $\Delta R^2 = 0.044$, (95% $CI = [0.034; 0.054]$).

Relative Deviation over Time ($RD[t]$) Figures 3.32 and 3.33 show the optimization history of each single replication using parameter schedules and no heuristic information. Because runtimes differed somewhat, the log-scaled x-axes use different scales. All cases portray searches as climbing steadily in solution quality, before replicating on the final solution for some time. Notably, only 1 of 1780 replications utilizing no schedule for the evaporation coefficient whatsoever, converged before reaching the abort criterion. Overall, 86 of the 162 conditions investigated here (53.086%) did not have a single replication converge before reaching the abort criterion, instead replicating on the final solution 256 times. As should be expected, the end scheduling had the greatest impact on the number of iterations after finding the final solution, with the number of ants achieving $\Delta R^2 = 0.365$, (95% $CI = [0.351; 0.379]$) and the evaporation coefficient ρ achieving $\Delta R^2 = 0.285$ (95% $CI = [0.269; 0.3]$). Scheduling K at the end of the search lead to more iterations being run after having found the final solution with fewer ants in each iteration, with an average of 231.227 being run for the early schedule, 188.082 for the late, and 166.139 for no schedule at all. Note that this does not imply that the absence of a schedule for K lead to quicker convergence. As described previously, quite the opposite is true in terms of absolute runtime. Regarding the schedule of ρ, earlier switches to lower values lead to less iterations after finding s^{gb} (137.698 for early schedules, 214.673 for late schedules, and 233.076 for no schedules).

In terms of the first half of the search, as indicated by the number of iterations required before finding the s^{gb}, the starting schedule for K was most important ($\Delta R^2 = 0.171$, 95% $CI = [0.038; 0.069]$). Here, using more iterations with fewer ants at the beginning of the search re-

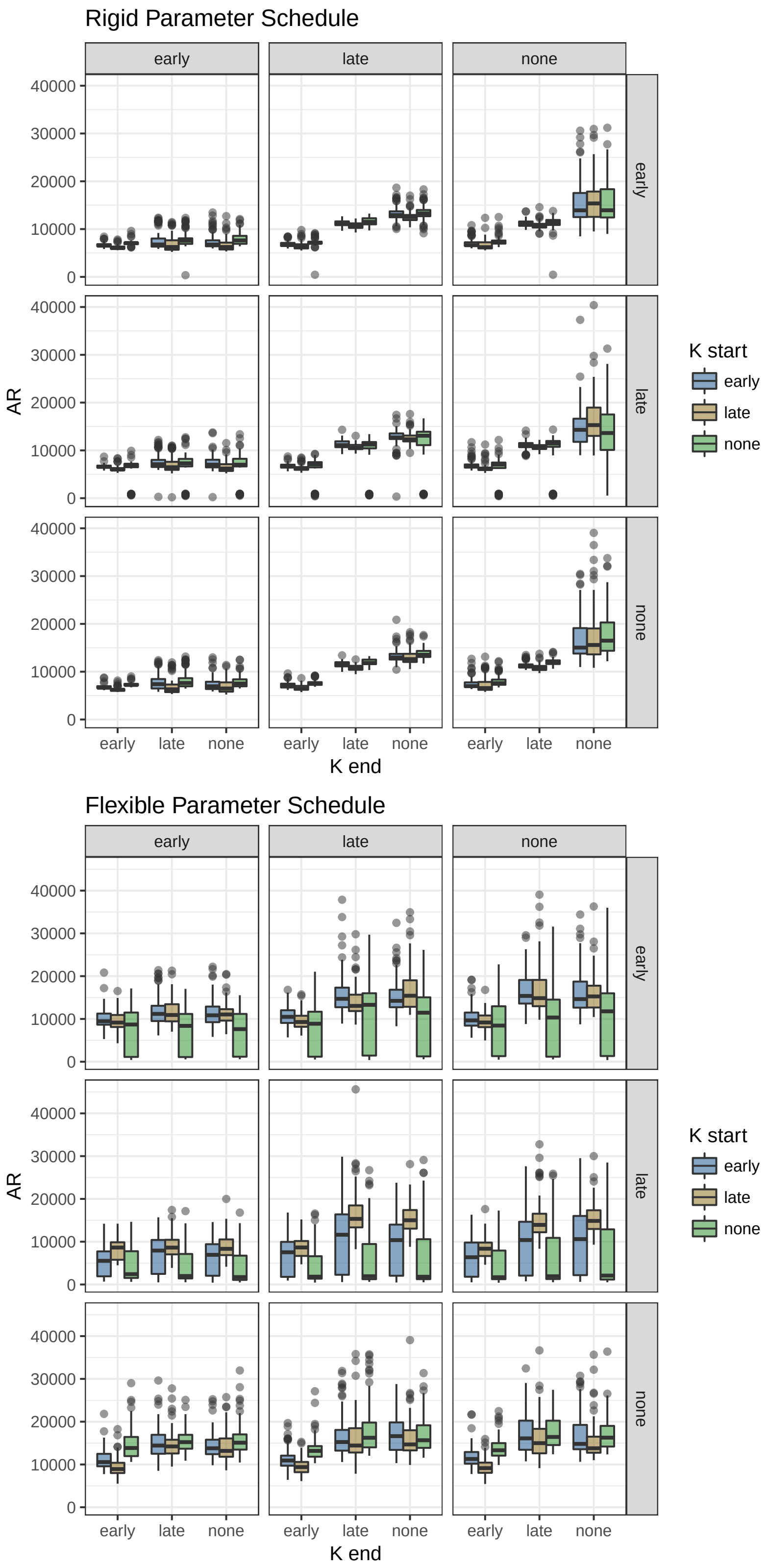


Figure 3.30: Boxplots of the absolute runtime (AR). Panels of the figure are for different combinations of ρ schedules. The starting schedules are depicted in rows, the ending schedules in columns.

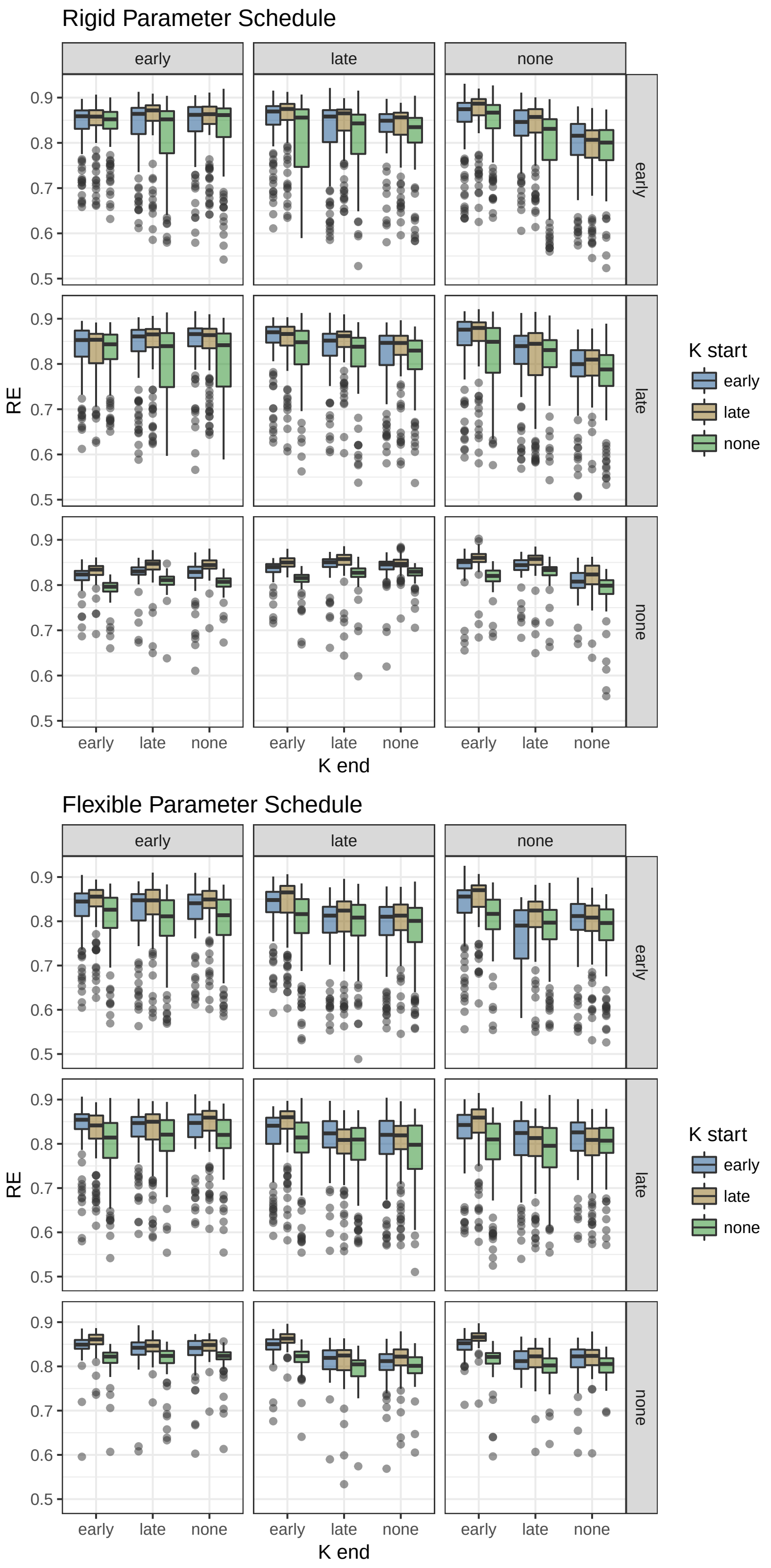


Figure 3.31: Boxplots of the relative exploration (RE). Panels of the figure are for different combinations of ρ schedules. The starting schedules are depicted in rows, the ending schedules in columns.

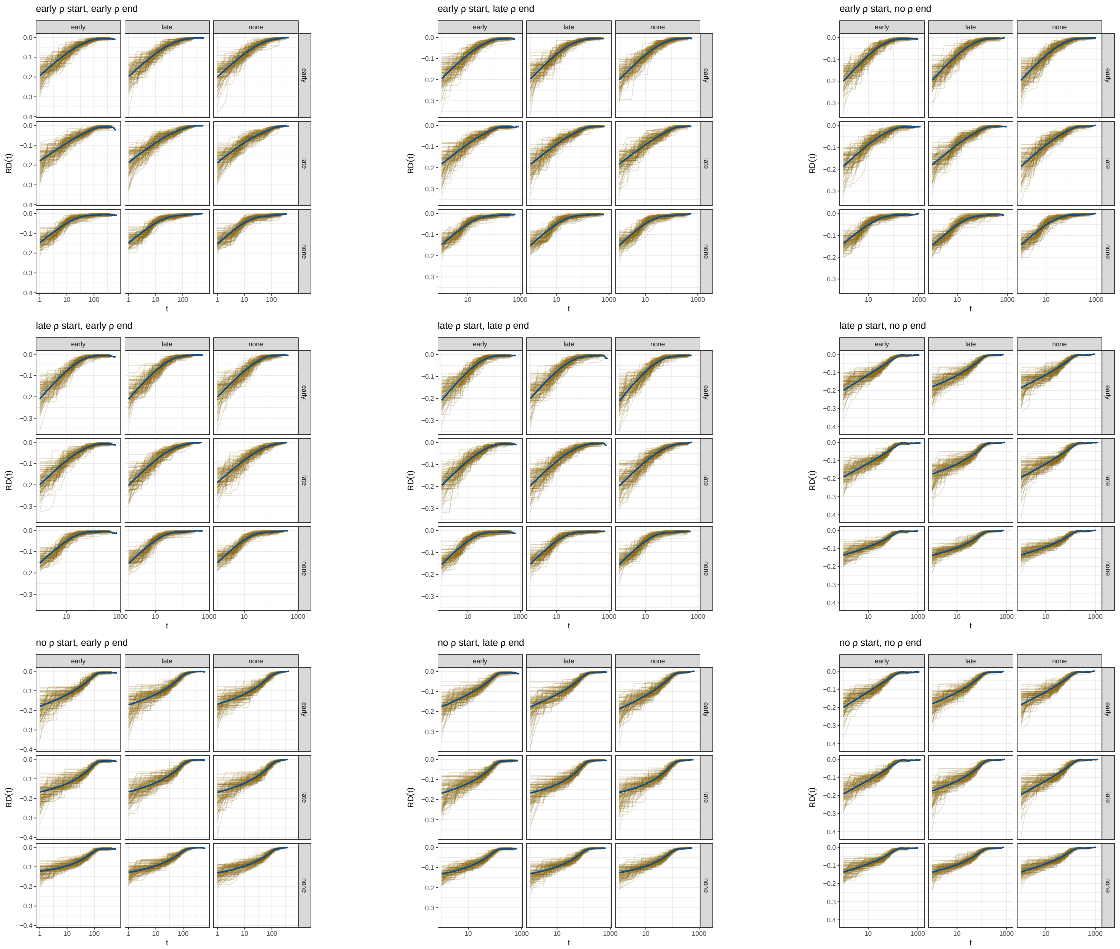


Figure 3.32: The relative deviation from $f(s^{opt})$ as a function of time $(RD[t])$ for the conditions using the rigid parameter schedule. Yellow lines represent the optimization history for each single replication, blue lines the LOESS-smoothed averages. The x-axis is log-scaled. Panels of each subfigure represent the K schedules, with starting schedules in rows and ending schedules in columns.

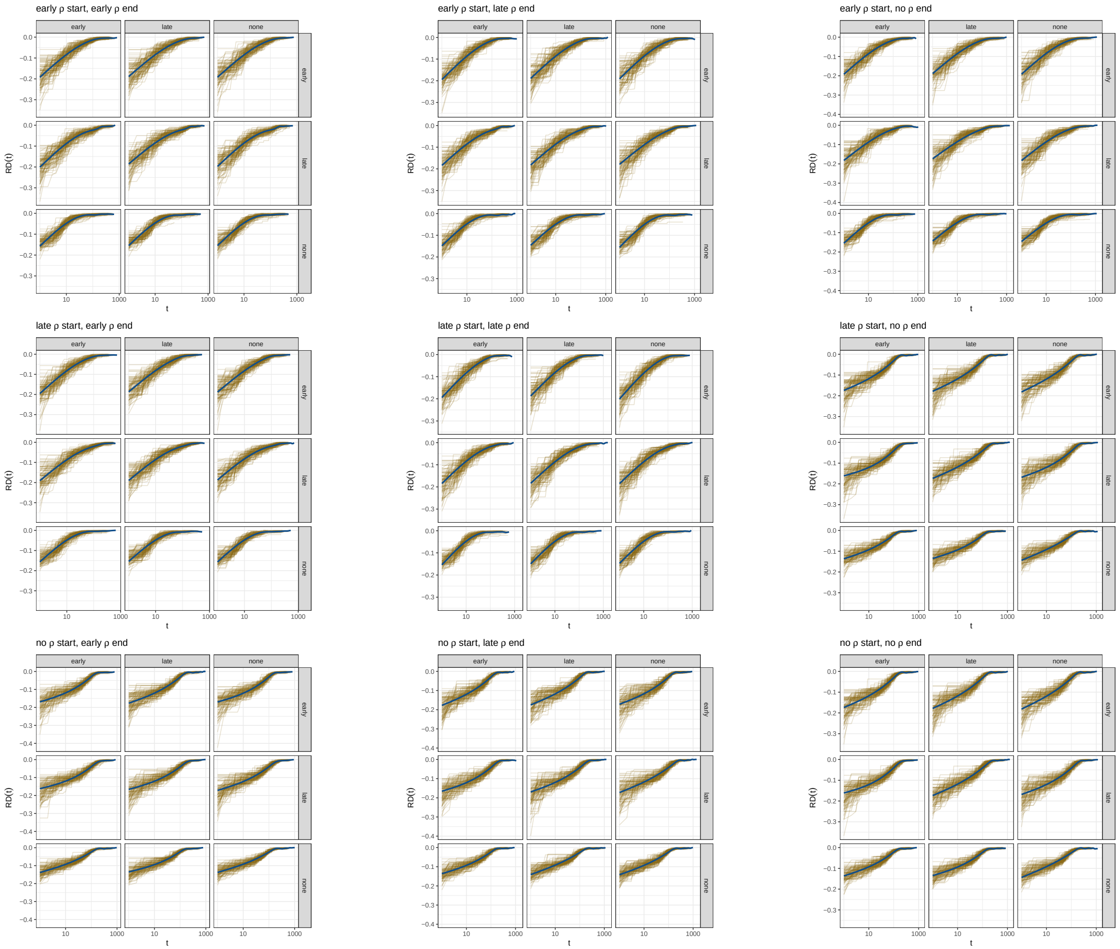


Figure 3.33: The relative deviation from $f(s^{opt})$ as a function of time $(RD[t])$ for the conditions using the flexible parameter schedule. Yellow lines represent the optimization history for each single replication, blue lines the LOESS-smoothed averages. The x-axis is log-scaled. Panels of each subfigure represent the K schedules, with starting schedules in rows and ending schedules in columns.

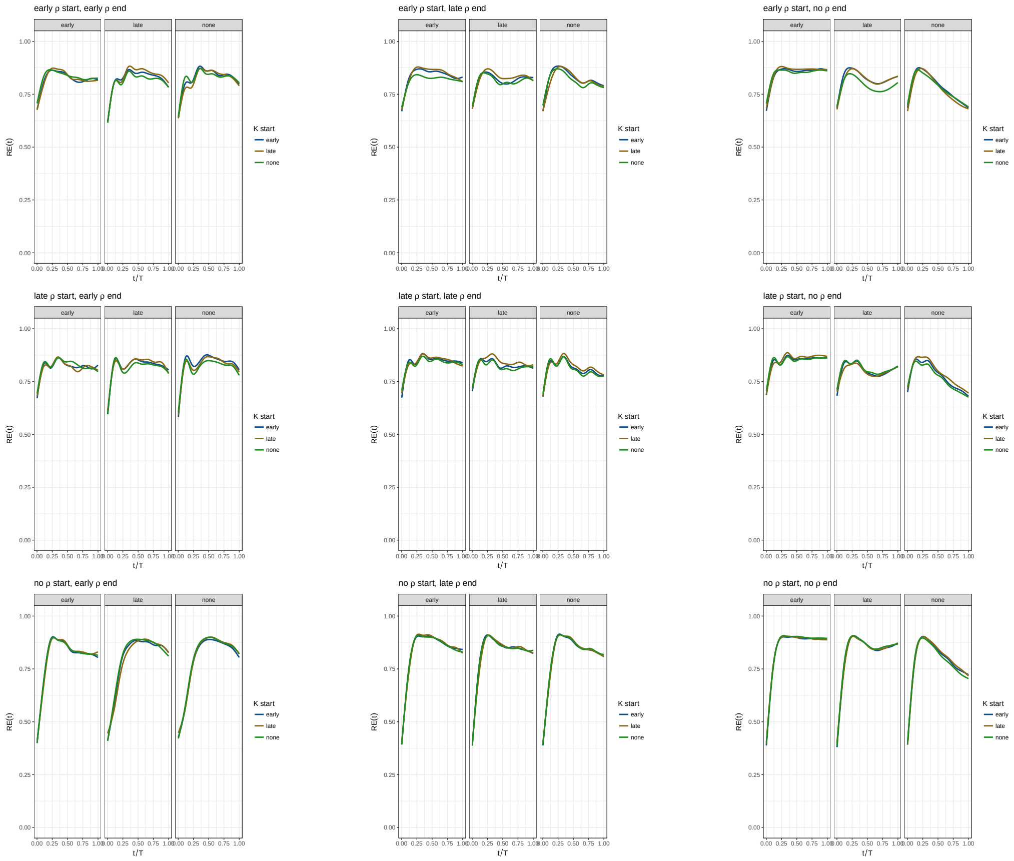


Figure 3.34: The relative exploration as a function of relative time (t/T) for the conditions using the rigid parameter schedule. Panels of each subfigure represent the ending K schedules in columns.

sulted in more iterations being required to find the best solution, with 183.063 iterations being the average for no ant schedule, 230.908 for the early ant schedule, and 268.952 for the late ant schedule. Again, this does not equate to longer runtimes in the beginning, because the ant schedules required less ants per iteration. In fact, for the rigid parameter schedule all the conditions required roughly 6250 CFAs before finding the s^{gb}.

The slightly convex curve of the LOESS smoothers in the lowermost parts of Figures 3.32 and 3.33 indicates less greedy improvement on the s^{gb} during the early phases of the search, when utilizing no schedule for ρ during this stage.

Relative Exploration over Time ($RE[t]$) Figure 3.34 shows the LOWESS lines of the $RE(t)$ for all conditions with a rigid parameter schedule and Figure 3.35 shows those of all conditions with a flexible parameter schedule. In both cases the most noticeable difference in $RE(t)$ is the difference in burn-in when comparing conditions with different starting schedules of ρ. Conditions

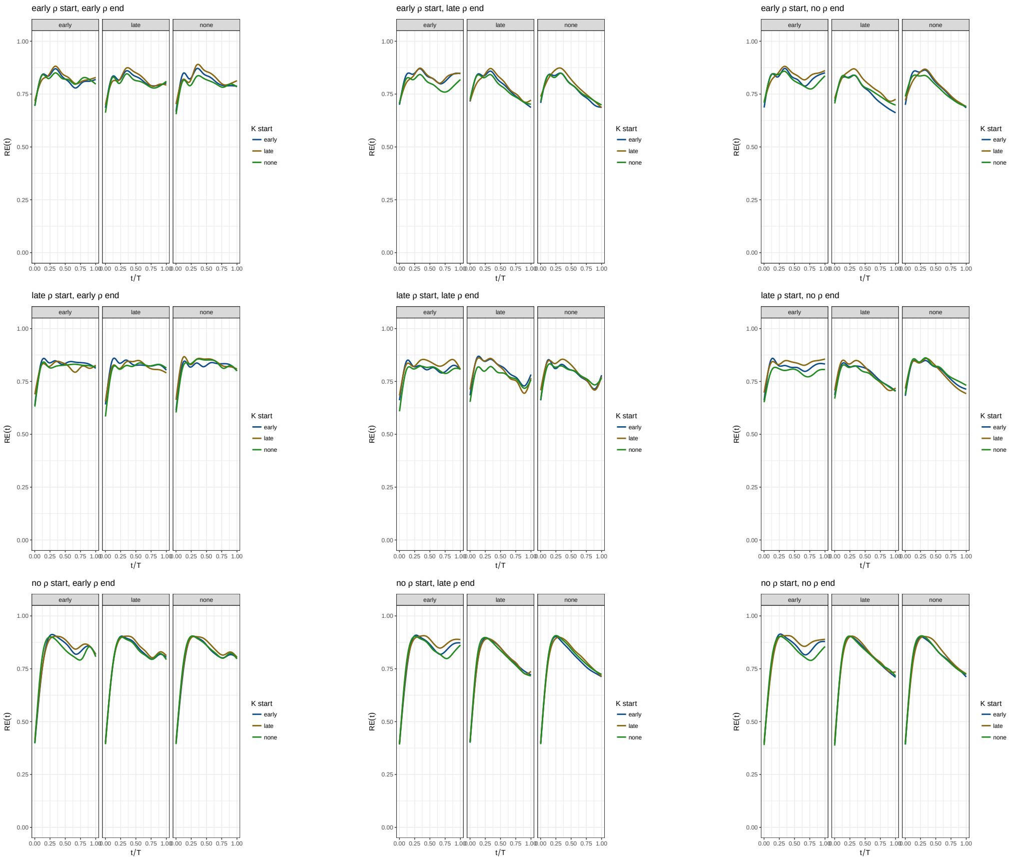


Figure 3.35: The relative exploration as a function of relative time (t/T) for the conditions using the rigid parameter schedule. Panels of each subfigure represent the ending K schedules in columns.

Table 3.20: Average number of iterations needed for the burn-in phase (T_{bi}) per condition with the average 90th percentile of RE in parentheses.

				early K end		late K end		no K end	
Schedule	ρ start	ρ end	K start						
rps	early	early	early	24.140	(0.992)	15.155	(0.982)	12.260	(0.978)
			late	15.515	(0.991)	10.949	(0.983)	10.800	(0.979)
			none	72.010	(0.997)	19.810	(0.997)	16.850	(0.995)
		late	early	19.455	(0.943)	13.596	(0.950)	11.340	(0.949)
			late	17.670	(0.935)	15.526	(0.961)	12.970	(0.930)
			none	55.790	(0.952)	42.420	(0.979)	12.630	(0.967)
		none	early	21.970	(0.934)	14.459	(0.928)	10.293	(0.913)
			late	11.918	(0.940)	11.525	(0.916)	10.316	(0.904)
			none	55.140	(0.943)	18.670	(0.944)	11.930	(0.929)
	late	early	early	27.917	(0.990)	11.434	(0.986)	14.141	(0.981)
			late	12.816	(0.990)	10.380	(0.987)	10.290	(0.982)
			none	64.680	(0.999)	17.960	(1.000)	13.930	(0.996)
		late	early	20.041	(0.950)	17.990	(0.969)	10.190	(0.952)
			late	11.740	(0.948)	10.768	(0.972)	10.051	(0.937)
			none	50.490	(0.954)	21.070	(0.979)	12.830	(0.975)
		none	early	13.283	(0.937)	13.667	(0.932)	9.160	(0.911)
			late	10.515	(0.943)	10.880	(0.926)	10.980	(0.921)
			none	36.690	(0.948)	19.050	(0.943)	11.910	(0.931)
	none	early	early	109.626	(0.992)	76.880	(0.960)	72.235	(0.977)
			late	101.293	(0.990)	70.697	(0.980)	67.255	(0.977)
			none	107.950	(0.996)	73.760	(0.996)	72.460	(0.997)
		late	early	110.480	(0.918)	84.768	(0.948)	68.960	(0.915)
			late	99.690	(0.923)	84.588	(0.940)	67.100	(0.932)
			none	103.020	(0.956)	88.770	(0.975)	72.080	(0.977)
		none	early	101.101	(0.925)	87.092	(0.917)	68.515	(0.902)
			late	93.596	(0.920)	86.545	(0.914)	64.212	(0.894)
			none	106.670	(0.946)	90.810	(0.945)	71.570	(0.926)
fps	early	early	early	11.290	(0.971)	10.120	(0.971)	10.640	(0.964)
			late	10.240	(0.979)	11.253	(0.956)	11.202	(0.964)
			none	20.700	(0.981)	12.610	(0.981)	16.500	(0.984)
		late	early	12.710	(0.929)	10.192	(0.913)	10.830	(0.899)
			late	10.182	(0.942)	9.680	(0.930)	10.062	(0.914)
			none	13.600	(0.939)	11.910	(0.931)	11.310	(0.933)
		none	early	10.566	(0.932)	8.653	(0.907)	11.283	(0.915)
			late	10.848	(0.942)	10.354	(0.920)	11.192	(0.921)
			none	16.900	(0.937)	11.810	(0.925)	11.390	(0.930)
	late	early	early	10.600	(0.982)	10.394	(0.970)	13.144	(0.977)
			late	11.969	(0.988)	11.160	(0.974)	10.495	(0.975)
			none	16.490	(0.994)	12.930	(0.999)	13.740	(0.997)
		late	early	10.296	(0.938)	11.180	(0.926)	11.130	(0.925)
			late	10.710	(0.945)	10.404	(0.935)	10.827	(0.929)
			none	15.320	(0.950)	12.320	(0.934)	13.160	(0.945)
		none	early	11.162	(0.945)	10.000	(0.927)	11.071	(0.922)
			late	11.190	(0.962)	10.469	(0.931)	10.909	(0.926)
			none	14.090	(0.953)	11.430	(0.940)	11.940	(0.946)
	none	early	early	83.439	(0.937)	63.590	(0.920)	64.408	(0.936)
			late	56.837	(0.938)	53.765	(0.924)	55.092	(0.916)
			none	78.620	(0.954)	71.390	(0.955)	73.000	(0.950)
		late	early	88.160	(0.910)	62.150	(0.899)	61.340	(0.899)
			late	59.293	(0.922)	55.720	(0.896)	53.919	(0.899)
			none	77.550	(0.934)	70.490	(0.929)	70.050	(0.931)
		none	early	82.576	(0.908)	61.000	(0.902)	62.121	(0.901)
			late	56.778	(0.931)	50.758	(0.903)	51.173	(0.917)
			none	75.550	(0.939)	68.920	(0.933)	71.750	(0.935)

utilizing any schedule on the evaporation coefficient lead to a much shorter burn-in, mainly due to constructing less improper solutions. On average, these conditions required 15.637 iterations before achieving their 90th percentile in $RE(t)$, while conditions without a starting schedule for ρ required 75.58. As noted previously, however, this does not negatively impact the overall RE of these conditions. The second-most visible effect in Figures 3.34 and 3.35 is that later scheduling of diminishing K at the end of the search decreases $RE(t)$ during those iterations. This is more visible in cases utilizing the fps, where this trend is also visible for late K schedules, while it is only observable for cases not utilizing any K schedule with the rps. The factor with the largest influence on overall $RE(t)$ - the K schedules at the beginning of the search - seems to have little influence on the form of the $RE(t)$ curves.

Summary In general, parameter schedules for situations without heuristic information had a substantial influence on runtimes and solution quality. Decreasing the number of ants during the later stages of the search lead to much shorter runtimes, but also resulted in the worst average performance regarding solution quality. Using ρ schedules during the beginning of the run resulted in much larger variability between different starts of the algorithm, leading to substantially worse worst-case performance in situations, in which ρ allowed for quick evaporation for longer periods. Constellations utilizing schedules for increases in K during the beginning and decreases in ρ during the end of the search performed best with regards solution quality, though they are among the slowest. In comparison to fixed parameter settings, these conditions resulted in better average and comparable worst-case performance without being substantially faster.

3.8.3 Scheduled Parameters with Sub-Optimal Heuristic Information

The results from Sections 3.7.3 and 3.8.2 show some general trends, which are combined in this section to derive the parameter settings used to evaluate the `stuart` approach with parameter schedules and sub-optimal heuristic information. In contrast to the parameter settings used in the previous section, the setting of α and β will be of crucial importance to the performance of the parameter schedules with heuristic information. As pointed out in Sections 3.1.2 and 3.1.3, these two non-linearity parameters greatly influence the speed and precision with which the algorithm moves through the search space. Additionally, Section 3.7.3 indicated a much less clear picture about the selection of a deposit rule, than was the case for conditions without heuristic information, necessitating an investigation of the search behavior with both the *ib* and the *gb* deposit. To prevent these three new variable parameters from increasing the number of evaluation conditions drastically, their variation is combined with only those conditions which showed the most promising results in the evaluation presented in Section 3.8.2 for parameter schedules with no heuristic information.

Regarding the first non-linearity coefficient, higher values of α generally lead to increasingly fast pheromone accumulation on promising choices. While $\alpha > 1$ is generally not recommended and Section 3.7.1 clearly showed these conditions to lead to sub-optimal solutions (albeit at a much faster pace), the evaluation of constant parameter settings with sub-optimal heuristic information revealed conditions with $\alpha = 1.5$ to be those with the highest success rates. However, conditions with $\alpha = 1$ exhibited a greater extent of exploration, leading to very similar average solution quality at the cost of - in cases with the iteration best deposit rule, drastically - longer runtimes. In this evaluation a parameter schedule beginning with $\alpha = 1$ and gradually increasing its value (to $\alpha = 1.5$ and then $\alpha = 2.5$) during the later stages of the search process is compared to conditions using only $\alpha = 1$. The latter is chosen as the constant reference setting, despite the slight advantages conditions with $\alpha = 1.5$ had in Section 3.7.3, because of the overwhelming amount of literature on $\mathcal{MM}$AS and similar strategies recommending that α not exceed 1. The increases in α are set at such late points in the search, to allow for exploration during the early phases and reinforce better choices during the later phases to concentrate search around these choices.

The opposite is done for the second non-linearity coefficient, β. As shown in Section 3.7.3, β has a dramatic influence on the tradeoff between selection quality (as measured by RD) and the selection consistency (SC), with larger values leading to solutions closer to the heuristically favored solution. To make use of this, β is gradually decreased during the early phases of the search in this evaluation, to ensure that heuristics guide the early search in the favored direction and become less important, once the search is within this subspace of possible solutions. As shown in Section 3.7.3 the degree of influence β has on the explorative behavior of `stuart` is dependent on the deposit rule, with the *gb* rule being less sensitive to heuristic information. Because of this, the condition with constant β differs, depending on the deposit rule with $\beta = 1.5$ for *ib* deposit and $\beta = 2.5$ for *gb* deposit.

The remaining two parameters which are scheduled in this evaluation, are the number of ants K and the evaporation coefficient ρ. The results shown in Section 3.8.2 indicate that scheduling the number of ants to reduce towards the end of the search process has a detrimental effect on the quality of the final solution. Thus, these conditions are not investigated in this section. On the other hand, beginning with low values of K and increasing this during the early search leads to better solution quality at the expense of increased runtime for conditions without heuristic information. To investigate the effect of the increase in cooperation between ants during the early search, K is scheduled during the beginning of the algorithm in line with the schedules used for the case without heuristic information.

While the schedule for values of K during the late search is detrimental to the algorithmic performance, the opposite is true for the evaporation coefficient ρ in the results shown in Section 3.8.2. In those cases, placing a schedule on ρ during the initial phases of the search lead to much higher variability in search results, often leading to very bad worst-case performance. Therefore schedules on ρ during the beginning of the search are excluded from this evaluation. On the other hand, decreasing ρ during the late stages of the search had no detrimental effect on

Table 3.21: Parameter schedules for the evaluation with sub-optimal heuristic information.

Parameter	Condition	Values		Schedule	
		ib	*gb*	*rps*	*fps*
α	early	(1,1.5,2.5)	(1,1.5,2.5)	(0,200,300)	(0,75,150)
	none	1	1	-	-
β	early	(2.5,1.5,1)	(2.5,1.5,1)	(0,30,60)	(0,30,60)
	none	1.5	2.5	-	-
K	early	(8,16,32)	(8,16,32)	(0,15,30)	(0,15,30)
	late	(8,16,32)	(8,16,32)	(0,30,60)	(0,30,60)
ρ	early	(.95,.8,.5)	(.95,.8,.5)	(0,150,200)	(0,75,150)
	late	(.95,.8,.5)	(.95,.8,.5)	(0,300,400)	(0,100,200)

solution quality but did decrease runtime in the evaluation for cases without heuristic information. Therefore, the evaporation coefficient is decreased during the late stages of the search in this evaluation, in line with the conditions shown in Section 3.8.1.

As pointed out above and in Section 3.7.3 for conditions with constant parameter settings, the deposit rule has substantial influence on the runtimes of the algorithm with heuristic information. To investigate this under conditions with scheduled parameter settings, both deposit rules are included in this evaluation. Finally, to determine the type of scheduling best suited for instances with heuristic information, both the *rps* and the *fps* are included in this evaluation.

Table 3.21 shows the evaluation design for conditions with scheduled parameters and sub-optimal heuristic information. The combinations result in a total of 64 conditions. As was the case in the preceding evaluations, each condition is replicated 100 times.

3.8.4 Results with Sub-Optimal Heuristic Information

Errors Across all conditions, 38 replications (0.594%) did not terminate normally. As was the case in all previous evaluations, the errors were due to the search not encountering a viable solution in the first iteration. Conditions did not differ in any respect during the first iteration, because all conditions used $K = 8$.

Success Rate (*SR*) Across all conditions the average $SR = 0.031$. Table 3.22 provides an overview of the condition specific SR. Of the five factors manipulated in this evaluation, two showed a substantial influence on the success rate: the non-linearity coefficient β ($\Delta R^2 = 0.253$, 95% $CI = [0.032; 0.473]$) and the deposit type ($\Delta R^2 = 0.208$, 95% $CI = [-0.015; 0.431]$). As is visible in Table 3.22, s^{opt} was recovered almost exclusively in conditions with a schedule for β. Beyond that, the *gb*-deposit rule resulted in a higher SR (0.003 vs. 0.058). Over all conditions, those combining a scheduled β with the *gb* deposit and a flexible parameter schedule resulted in the highest success rates, irrespective of the scheduling chosen for α and ρ.

Table 3.22: Condition-specific SR for cases with scheduled parameters and sub-optimal heuristic information.

			ib Deposit				*gb* Deposit			
			early K		late K		early K		late K	
	α	β	early ρ	late ρ	early ρ	late ρ	early ρ	late ρ	early ρ	late ρ
rps	early	early	0.000	0.000	0.000	0.000	0.010	0.121	0.010	0.170
		none	0.010	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	none	early	0.030	0.000	0.010	0.000	0.010	0.060	0.030	0.121
		none	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
fps	early	early	0.000	0.020	0.000	0.010	0.130	0.162	0.202	0.162
		none	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	none	early	0.020	0.000	0.000	0.000	0.121	0.202	0.162	0.202
		none	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Relative Deviation (RD) The average RD across all conditions was -0.016, indicating that the average solution in this evaluation was 1.629% worse in terms of the quality function than the optimal solution. None of the replications resulted in an RD worse than that of the heuristically favored solution. In fact, the worst single $RD_{s^{gb}} = -0.064$, where $RD_{s^{med}} = -0.238$.

Figure 3.36 shows the boxplots of the RD for conditions with parameter schedules and sub-optimal heuristic information. As was the case for the SR, β had the largest effect on the RD with a $\Delta R^2 = 0.356$ (95% $CI = [0.334; 0.379]$). As indicated by Figure 3.36, conditions with a fixed setting for β performed notably worse on average. This effect was magnified when using the global-best deposit-rule, where the difference between the settings for β (Cohen's $d = 1.953$, 95% $CI = [1.869; 2.038]$) was even more pronounced than with the iteration-best deposit (Cohen's $d = 1.201$, 95% $CI = [1.125; 1.276]$). Of the remaining factors, only the deposit rule had a detectable, albeit small, impact on the RD ($\Delta R^2 = 0.042$, 95% $CI = [0.016; 0.068]$), with the iteration-best deposit rule achieving slightly better average RD values ($RD_{ib} = -0.014$, $RD_{gb} = -0.018$).

Ranking all conditions with regard to RD shows that the 31 best-performing conditions in terms of RD all utilized a schedule for β. Table 3.23 shows the five conditions which performed best, either on average or in the worst-case, with regards to RD. While the case is clear with regards to settings for β, other parameters showed more inconsistent performance. Generally, however, there was a tendency for the *gb*-deposit to perform better on average, while the *ib*-deposit performed better in the worst case, showing slightly lower variability across replications ($sd_{ib} = 6.7 \times 10^{-3}$, $sd_{gb} = 1.15 \times 10^{-2}$,).

Selection Consistency (SC) The average selection consistency across all replications was 0.483, indicating that roughly 9 items were shared between any s^{gb} and the heuristically favored solution. Recall, that the SC of s^{opt} is 0.389. Figure 3.37 shows the condition specific SC. As

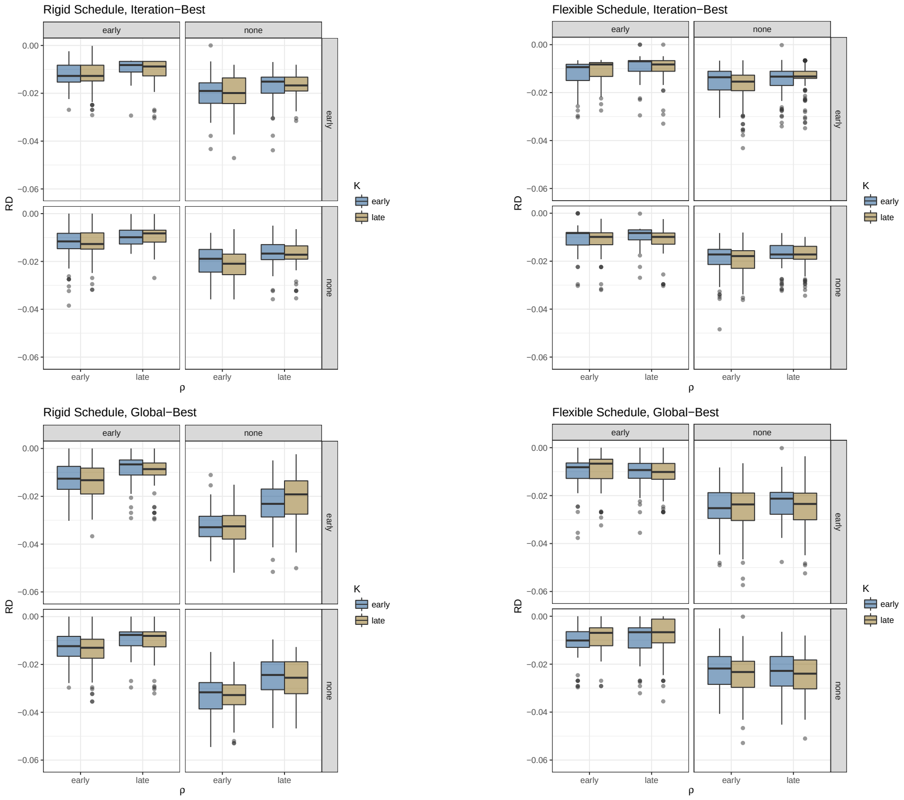


Figure 3.36: Boxplots of the relative deviation (RD) of the final solutions from the optimal solution. Panels represent different values for α in rows and β in columns.

Table 3.23: Average and worst-case RD for the five best performing conditions in each respect. The ranks pertain to the total of all 64 conditions with sub-optimal heuristic information.

						Average		Worst-Case			
		α	β	ρ	K	RD	Rank	RD	Rank	SR	AR
gb	*fps*	none	early	late	late	-0.008	1	-0.036	34	0.202	8548.283
gb	*rps*	early	early	late	early	-0.008	2	-0.029	7	0.121	8360.404
gb	*fps*	none	early	early	late	-0.009	3	-0.029	7	0.162	8604.848
gb	*fps*	early	early	early	late	-0.009	4	-0.032	28	0.202	8280.323
ib	*rps*	early	early	late	early	-0.009	5	-0.029	10	0.000	9434.560
ib	*fps*	none	early	late	early	-0.010	14	-0.027	2	0.000	18564.160
ib	*rps*	none	early	late	early	-0.010	15	-0.017	1	0.000	13967.360
gb	*fps*	early	early	late	late	-0.010	16	-0.027	2	0.162	8258.909
ib	*rps*	none	early	late	late	-0.010	17	-0.027	2	0.000	13383.040
ib	*rps*	early	early	early	early	-0.013	27	-0.027	2	0.000	6088.081

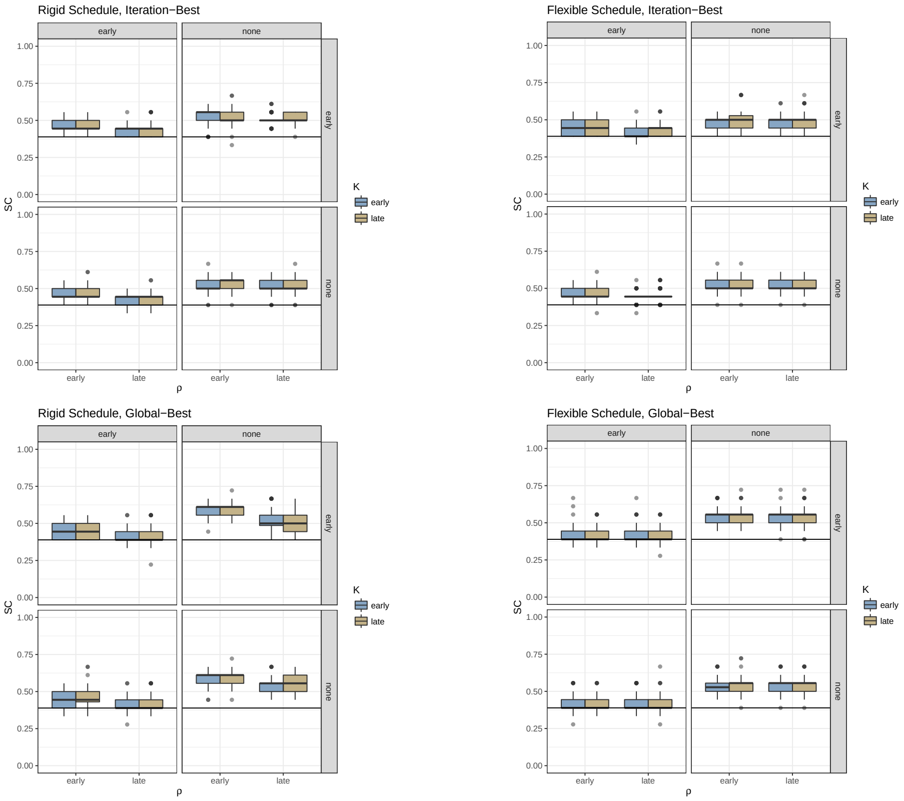


Figure 3.37: Boxplots of the selection consistency (SC) of the conditions with sub-optimal heuristics and scheduled parameter settings. Panels represent different values for α in rows and β in columns. The horizontal line depicts the selection consistency of the optimal solution.

this figure indicates, the scheduling of β was the most important factor for selection consistency with a $\Delta R^2 = 0.356$ (95% $CI = [0.334; 0.379]$). Those conditions scheduling β achieved a lower SC, with the effects differing somewhat, depending on the combination with either the gb (Cohen's $d = -2.199$, 95% $CI = [-2.287; -2.111]$) or the ib deposit-rule (Cohen's $d = -1.353$, 95% $CI = [-1.43; -1.276]$). However, the effect of the deposit rule on SC was minimal, as was that for all other factors, with none of the 95% confidence intervals of the ΔR^2 excluding zero.

Absolute Runtime (AR) Across all 6400 replications, the average AR was 9585.475, with the slowest single replication requiring at total of 33784 CFAs to be estimated. Figure 3.38 shows the boxplots of the AR. The most visible effect is that of the schedule type on the variability of runtimes, with the rps resulting in much more homogeneous AR across replications ($SD_{rps} = 3174.164$, $SD_{fps} = 4311.913$). The schedule type also had the second largest effect on the average AR, with a $\Delta R^2 = 0.193$ (95% $CI = [0.169; 0.218]$).

The largest determinant of absolute runtime was the deposit rule ($\Delta R^2 = 0.195$, 95% $CI =$

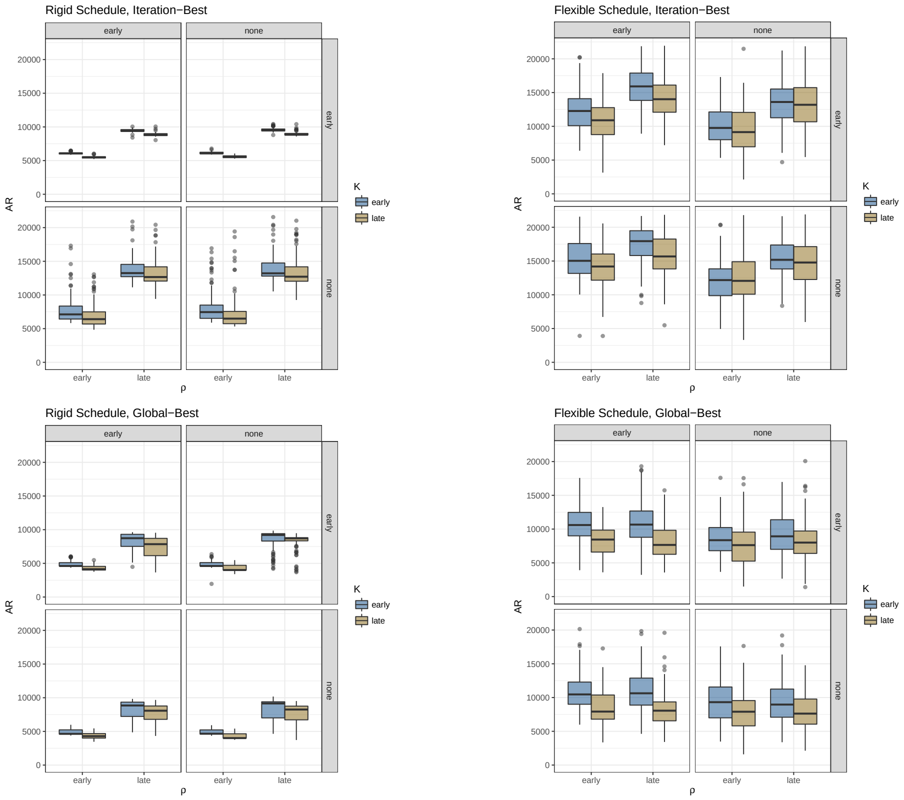


Figure 3.38: Boxplots of the absolute runtime (AR) of the conditions with sub-optimal heuristics and scheduled parameter settings. Panels represent different values for α in rows and β in columns.

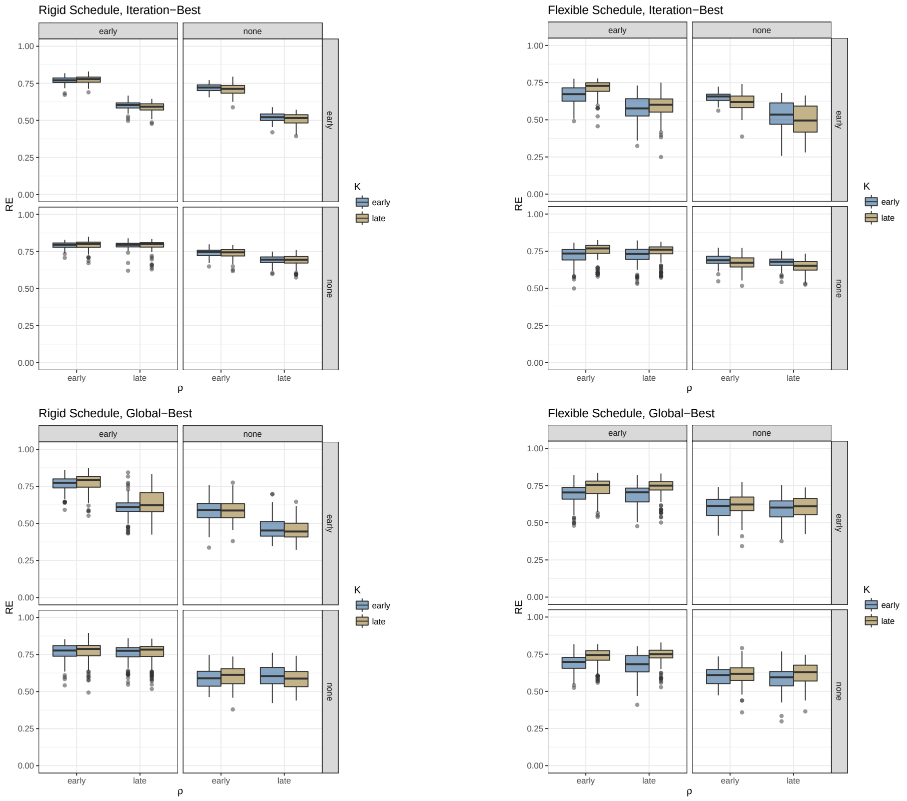


Figure 3.39: Boxplots of the relative exploration (RE) of the conditions with sub-optimal heuristics and scheduled parameter settings. Panels represent different values for α in rows and β in columns.

$[0.17; 0.22]$) with ib conditions requiring much longer runtimes ($AR_{ib} = 11444.184$) than their gb counterparts ($AR_{gb} = 7725.597$). Additionally, the scheduling of ρ had a substantial impact on AR ($\Delta R^2 = 0.12$, 95% $CI = [0.095; 0.144]$), with the earlier schedule leading to shorter runtimes. Of the non-linearity coefficients, scheduling α had a larger impact on AR ($\Delta R^2 = 0.03$, 95% $CI = [0.006; 0.053]$) than did schedules of β ($\Delta R^2 = 0.007$, 95% $CI = [-0.016; 0.03]$), but both had close to no impact on the runtimes at all. The same is true for the scheduling of K, which did not have a substantial impact on AR ($\Delta R^2 = 0.015$, 95% $CI = [-0.008; 0.038]$).

The eight conditions with the shortest runtimes all utilized early schedules of the evaporation coefficient, the gb deposit, and the rps. They differed only in in the scheduling of K, with the first four using the late switch to larger numbers of ants and the latter four all utilizing the early switch. On the opposite end, the seven slowest conditions all used the ib deposit-rule and the fps. Additionally, six of those seven used the late schedule in ρ.

Relation Exploration (RE) Across all conditions, the average relative exploration was 0.663, indicating that roughly two thirds of all solutions that were generated, were viable and unique within that replication. Figure 3.39 shows the boxplots of the RE across all conditions. The non-linearity coefficients were the most important determinants of the RE, with β achieving a $\Delta R^2 = 0.223$ (95% $CI = [0.197; 0.248]$) and α accomplishing a $\Delta R^2 = 0.1$ (95% $CI = [0.074; 0.127]$). In the case of scheduling β, not utilizing a schedule lead to less exploration (0.713 vs. 0.613), while the opposite was true for α (0.63 vs. 0.697).

Scheduling the evaporation coefficient ρ had the third largest impact on the RE ($\Delta R^2 = 0.085$, 95% $CI = [0.059; 0.112]$). As is visible in Figure 3.39, this effect is dependent on the combination with other parameter settings. The impact of the two schedules of ρ on RE was most prominent in combination with the schedules on α: the combinations of an early schedule on ρ, a scheduled α (RE = 0.685), an early schedule on ρ, no schedule on α (RE = 0.703), or a late schedule on ρ and no schedule on α (RE = 0.691) all resulted in extremely similar relative exploration. Combining a schedule on α with a late schedule on ρ, however, decreased the RE noticeably (RE = 0.574).

Of the conditions with the lowest RE, the worst six all used combinations of the α schedule and no β schedule. RE was below .5 for three of these. On the other end of the spectrum, the four conditions with the highest RE all combined no schedule on α, a scheduled β, the iteration-best deposit-rule, and the *rps*, resulting in an RE just shy of .8.

Relative Deviation over Time ($RD[t]$) Figures 3.40 and 3.41 show the optimization history of all replications using parameter schedules and sub-optimal heuristic information. The x-axes are log-scaled to better accommodate the differences in runtimes.

On average, replications required 288.887 iterations before finding the s^{gb}. The most important determinant for this number was the schedule type ($\Delta R^2 = 0.262$, 95% $CI = [0.237; 0.287]$), with the flexible parameter schedule requiring significantly more iterations before finding the final solution (227.419 vs. 350.433). Beyond this, the deposit-rule had a noticeable impact on the number of iterations spent in this phase ($\Delta R^2 = 0.16$, 95% $CI = [0.134; 0.185]$). In this case, conditions using the *ib* deposit-rule needed 336.915 iterations before finding the final solution, while conditions with the *gb* deposit-rule needed only 240.83. The schedule chosen for K was the only remaining factor with a detectable influence on this number of iterations ($\Delta R^2 = 0.16$, 95% $CI = [0.134; 0.185]$). Overall, conditions with a late switch to higher values of K used more iterations to settle on a final solution (284.17 vs. 293.593). However, this effect was negligible in conditions using the *rps* (226.822 vs. 228.014), while it was detectable in conditions with the *fps* (341.554 vs. 359.297). Keep in mind, that more iterations does not necessarily imply longer runtimes for schedules with K, because later switches to higher values also mean that more iterations use less ants, thereby balancing the number of total CFAs estimated.

After finding a final solution, it was replicated as best an average of 58.372 times. This is

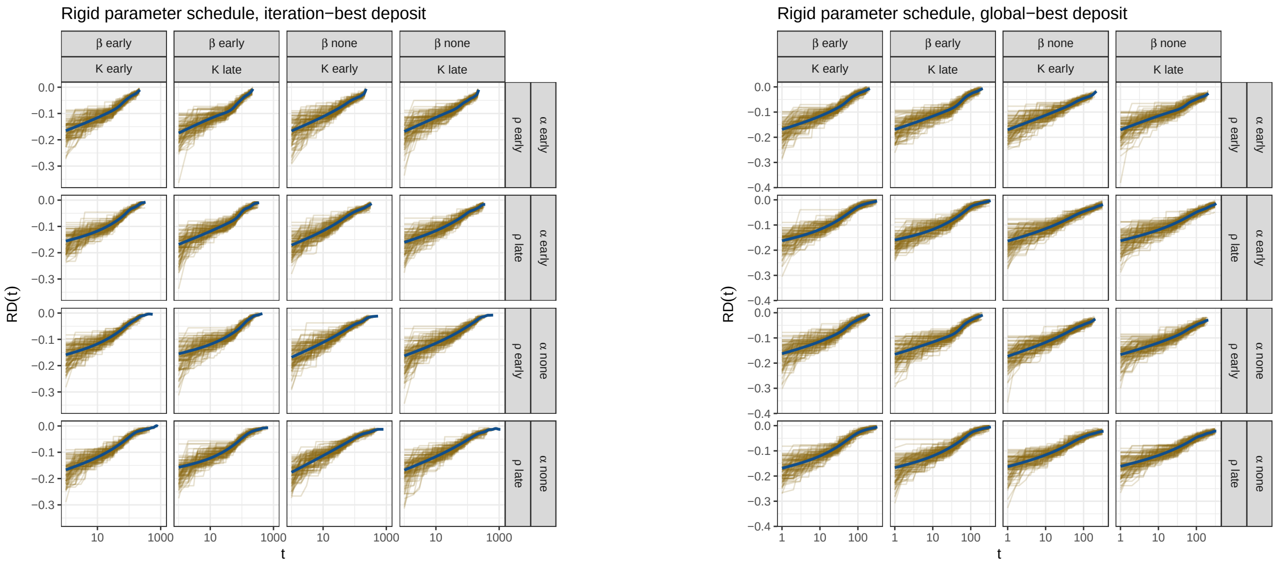


Figure 3.40: The relative deviation from $f(s^{opt})$ as a function of time ($RD[t]$) for the conditions using the rigid parameter schedule. Yellow lines represent the optimization history for each single replication, blue lines the LOESS-smoothed averages. The x-axis is log-scaled. The two sub-figures represent the different deposit rules.

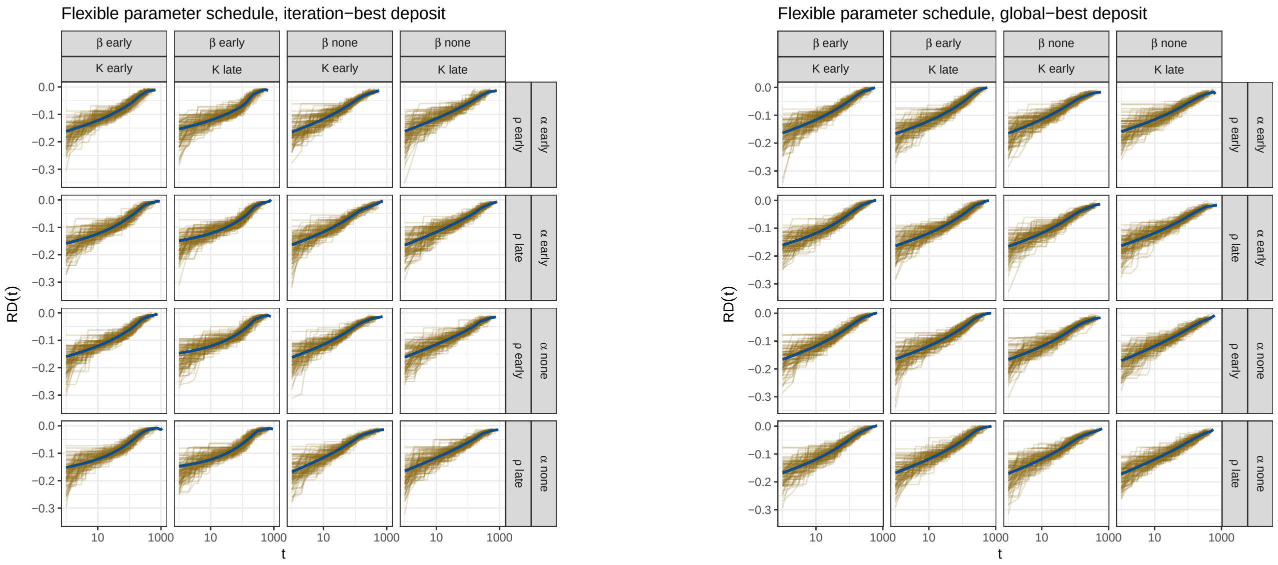


Figure 3.41: The relative deviation from $f(s^{opt})$ as a function of time $(RD[t])$ for the conditions using the flexible parameter schedule. Yellow lines represent the optimization history for each single replication, blue lines the LOESS-smoothed averages. The x-axis is log-scaled. The two sub-figures represent the different deposit rules.

visible in Figures 3.40 and 3.41 in that the overwhelming proportion of the optimization history showed increases in $RD(t)$. In fact, of the 6362 replications which did not result in errors, only 27 (0.424%) reached the abort criterion, while all others converged in accordance to the criteria set by Equation (2.11). The parameter schedule type was especially important in determining the number of iterations replicating s^{gb} ($\Delta R^2 = 0.138$, 95% $CI = [0.113; 0.163]$), with the fps, again, resulting in more iterations. Beyond the schedule type, the schedule imposed on the evaporation coefficient ρ influenced the length of the late search ($\Delta R^2 = 0.09$, 95% $CI = [0.064; 0.116]$), with early switches to lower values of ρ predictably leading to less iterations spent in this phase. The schedule of the non-linearity coefficient α was the final determinant of the number of iterations replicating the s^{gb} ($\Delta R^2 = 0.068$, 95% $CI = [0.041; 0.094]$), with a scheduled α leading to less iterations spent in replications of the s^{gb}.

Relative Exploration over Time ($RE[t]$) Figure 3.42 shows the LOWESS lines across all replications of each condition for the $RE(t)$. The most notable effect is the one the schedule placed on α has on relative exploration. Especially when combined with late changes to ρ, relative exploration was extremely low during the later stages of the search in conditions utilizing a schedule on α. Interestingly, conditions with the rps were more prone to this effect, whereas conditions utilizing the fps and the gb deposit rule did not portray this phenomenon. Another peculiarity is that, in conditions with the global-best deposit-rule, the initial burn-in was followed by a short peak and a subsequent valley in $RE(t)$. This specific course was much less pronounced for conditions utilizing the ib deposit.

Table 3.24 shows the number of iterations required to reach the 90th percentile in $RE(t)$ and its associated $RE(t)$-value. Generally, utilizing a schedule on β lead to a higher 90th percentile of the $RE(t)$ (0.87 with a scheduled β, 0.801 with a fixed β), which was reached after more iterations (27.846 vs. 16.875). Similar patterns are observable for the schedule type (where the rps reached higher values slower) and the deposit rule (where the same was true for the ib deposit). However, differences in the actual values of the 90th percentiles of $RE(t)$ were minute, which is line with the earlier findings concerning the total RE. Note that schedules on α had little to no impact on the values depicted in Table 3.24, because they indicate early exploration. The schedules for α only applied during the later stages of the search, making their impact on the explorative behavior visible in Figure 3.42, but not Table 3.24.

Summary In general, utilizing parameter schedules can have a substantial impact on runtimes when utilizing sub-optimal heuristics. Both deposit rules resulted in extremely similar solutions in terms of RD and SC, but the gb deposit rule proved substantially faster - mostly due to a quicker early search, as indicated by the $RD(t)$. The non-linearity coefficient β, scaling the influence of the heuristic information, had the largest overall impact on a vast array of performance measures. In terms of Success Rate as well as Relative Deviation, utilizing a schedule for β resulted in

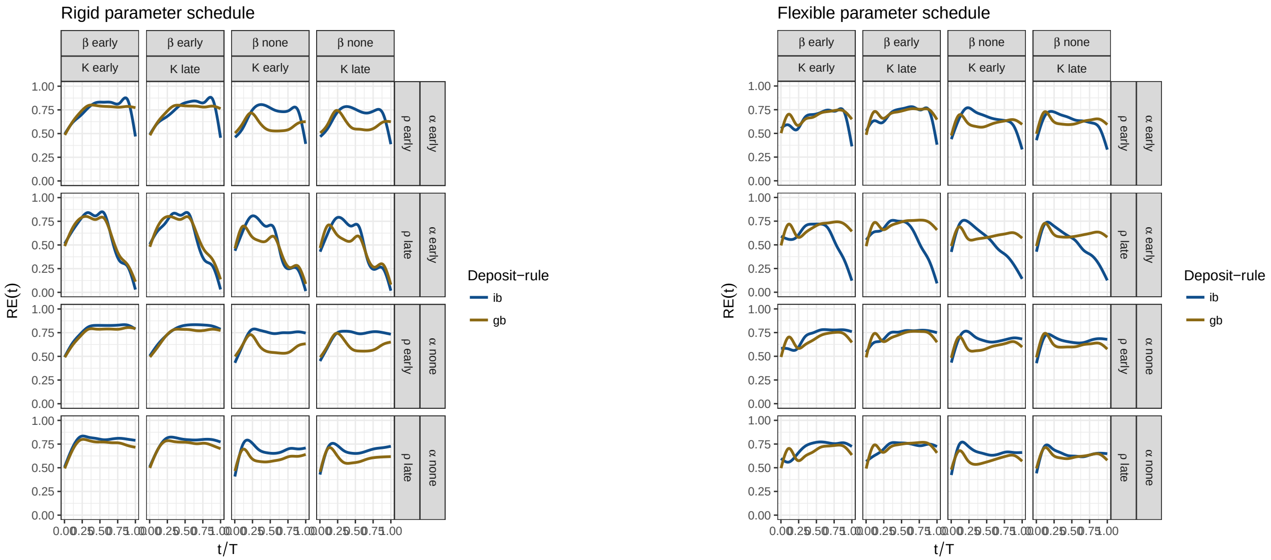


Figure 3.42: The relative exploration as a function of relative time (t/T) for the conditions using the rigid parameter schedule.

Table 3.24: Average number of iterations needed for the burn-in phase (T_{bi}) per condition with the average 90th percentile of RE in parentheses.

				early K				late K			
Schedule	Deposit	α	β	early ρ		late ρ		early ρ		late ρ	
rps	ib	early	early	49.939	(0.902)	42.390	(0.906)	31.340	(0.876)	23.260	(0.871)
			none	41.610	(0.857)	23.430	(0.849)	24.439	(0.813)	18.485	(0.812)
		none	early	62.360	(0.911)	45.646	(0.908)	61.770	(0.901)	45.080	(0.896)
			none	36.424	(0.870)	26.950	(0.865)	33.232	(0.837)	19.450	(0.831)
	gb	early	early	28.091	(0.883)	27.737	(0.891)	26.626	(0.858)	20.410	(0.867)
			none	10.337	(0.764)	9.152	(0.791)	6.670	(0.721)	6.434	(0.737)
		none	early	32.100	(0.888)	26.900	(0.886)	30.350	(0.876)	23.111	(0.877)
			none	9.070	(0.766)	11.180	(0.795)	12.867	(0.760)	9.620	(0.772)
fps	ib	early	early	23.550	(0.830)	13.910	(0.852)	16.657	(0.820)	16.500	(0.837)
			none	21.980	(0.811)	16.475	(0.805)	21.357	(0.791)	15.430	(0.777)
		none	early	27.727	(0.864)	32.280	(0.867)	23.060	(0.859)	18.400	(0.857)
			none	23.657	(0.824)	20.786	(0.819)	21.370	(0.810)	16.622	(0.806)
	gb	early	early	15.280	(0.838)	16.768	(0.872)	17.333	(0.840)	20.899	(0.876)
			none	9.480	(0.781)	10.150	(0.816)	9.071	(0.763)	10.260	(0.804)
		none	early	15.980	(0.837)	16.566	(0.875)	18.212	(0.834)	20.848	(0.879)
			none	12.200	(0.782)	11.980	(0.820)	9.459	(0.760)	10.380	(0.811)

much better solutions, while those utilizing a constant setting for β were more consistent with the heuristically favored solution and showed less relative exploration. Overall, the scheduling type had little influence on RD and SR, but the fps required longer and less consistent runtimes to converge. However, those conditions resulting in the best performance in terms of RD either utilized the fps or combined the rps with a late schedule for ρ, resulting in similar performance at similar runtimes. The schedule imposed on the number of ants had little effect whatsoever.

3.9 Discussion of the Evaluation Results

The first core result of the evaluation is that, given adequate parameter settings, the `stuart` approach can be used to select items from a pool to create a scale (or short-scale) which fulfills the criteria established in the pheromone function. In some cases, without requiring additional information about the item pool (i.e. in cases without heuristic information), the approach was able to construct the *one* optimal of 334569553920 possible solutions 48% of the time. Even when failing to find the optimal solution, the constructed solutions were often not much worse than optimal in terms of the pre-defined quality function. This behavior is, however, dependent upon using appropriate parameter settings. While some select, instance-specific settings lead to optimal behavior of the `stuart` approach, only few parameter constellations will result in truly bad search results. The previous sections of this chapter provided an overview of the parameters in the `stuart` approach and their expected general influence on the behavior of the search algorithm (Section 3.1), an in-depth application of the approach to the case of item selection from the Ryff-Scale (Sections 3.2 to 3.4), as well as a comprehensive evaluation study of the influence of parameter settings on the performance of the `stuart` approach in this application (Sections 3.5 to 3.8). This section is intended to integrate the overwhelming amount of information generated by those sections, discuss the general influence of the various parameters on the performance of the `stuart` approach, and close with some general recommendations regarding parameter settings in applications.

Heuristic Information (H) Regarding heuristic information, results of the evaluation are somewhat unclear. Section 3.6 provides the results regarding constant parameter settings for the three specific instances where no heuristic information is provided (Section 3.7.1), heuristics bias the search towards the optimal solution (Section 3.7.2), and heuristics bias the search towards a sub-optimal solution (Section 3.7.3). Generally, as should be expected, results are best when heuristics already provide the correct solution. The much more relevant conclusion from the overall results regarding the different heuristic information is, however, that providing heuristics can make the performance of `stuart` substantially worse. While the best performing conditions without heuristic information lead to an average Relative Deviation of -0.002, conditions supplied with sub-optimal heuristics were not always able to overcome the “false information” provided and resulted in an $RD = -0.007$ at the best of times.

Deposit Rule Providing heuristic information changes which deposit rule should be favored. In cases without heuristic information, the iteration-best deposit-rule consistently outperforms the rule in which only the global-best solutions deposit pheromones, while the exact opposite is true for conditions in which heuristic information is provided. This is especially true when considering the absolute runtimes in addition to the the quality of the final solutions. The search is always quicker when using the global-best deposit rule, which is unsurprising given the fact that there is less variability in which pheromones increase and evaporate after each iteration. In cases in which the selection process is biased by $\mathbf{H}$, the difference in runtimes becomes more drastic - without heuristics the ratio of average AR of *ib* vs. *gb* conditions is roughly 2.03, with sub-optimal heuristics it is approximately 2.365. However, the longer runtime is accompanied by an increase in exploration, which allows the *ib* deposit-rule to outperform its *gb* counterpart when the selection is not biased by heuristics. In cases where sub-optimal heuristic information is provided, the *gb* deposit-rule proves less influenced by the search bias which is introduced (as indicated by the lower selection consistency with the heurisitcally favored solution) and manages to outperform the *ib* deposit rule in terms of solutions and runtime. However, it should be noted, that when using very good heuristic information, this relative robustness of the *gb* deposit rule may also be detrimental, because it does not allow the search to be steered towards promising areas of the search space as easily. As pointed out in Section 3.8.3, this can be assuaged by increasing the value of β, which increases selection consistency without increasing runtimes for situations in which the *gb* deposit is used.

Non-linearity of pheromones (α) Regarding α, results mostly conform to prior studies, whereby values of $\alpha > 1$ lead to overly fast convergence to lower-value solutions (Alaya et al., 2004; Favaretto et al., 2009; Stützle, 1998; Stützle et al., 2010; Wei, 2014). Interestingly, this is not globally true for cases with sub-optimal heuristics. In cases with constant parameters, $\alpha = 1$ and $\alpha = 1.5$ show extremely similar results when combined with the most promising settings for the remaining parameters. The same is true for schedules which increase α during the late stages of the search to facilitate convergence. It should be noted that this did not, in fact, speed up convergence. However, when using either optimal or no heuristics, conditions with $\alpha = 1$ lead to the best performance in terms of quality. Due to this finding, there is only very limited foundation to recommending any value other than $\alpha = 1$.

Non-linearity of heuristic information (β) The non-linearity coefficient β has the predictable effect of increasing the influence of the heuristic information on the search and the final solution. This means that, in situations in which optimal heuristics are provided, runtimes are shorter and solution quality is higher, the larger the value of β. On the other hand, increasing

β in situations with sub-optimal heuristics will result in worse solutions without any benefit to runtimes. However, results from the conditions with scheduled parameter settings indicate that imposing a schedule on β during the early stages of the search, may provide benefits to overall performance, albeit without decreasing AR substantially. Specifically, scheduling β may prove beneficial in situations in which the early search is plagued by a substantial proportion of solutions being improper. Results indicate that the burn-in using a schedule on β is comparable to situations utilizing high values throughout the entire search, while simultaneously resulting in much better RD and SR values.

Evaporation coefficient (ρ) Predictably, lower static values of the evaporation ρ lead to shorter runtimes (because the disappearance of bad choices is accelerated). However, this comes at the cost of worse and substantially more variable solution quality. In situations without heuristic information and constant parameter settings, however, the best-performing conditions were some with $\rho = .8$, indicating a very specific "sweet spot" for the balance between forgetting bad choices and enough iterations for exploration. However, introducing sub-optimal heuristics shifts this trend towards higher values of ρ, most likely due to the necessity of generating more solutions to overcome the misleading directions provided by these heuristics. Scheduling ρ to increase during the early stages of the search has no substantial benefit with the added detriment of increasing solutions variability, thus decreasing worst-case performance. However, scheduling a decrease in ρ during the later stages of the algorithm provides a potential reduction in runtime, with practically no impact on the overall solution quality.

Number of ants (K) The number of ants has a substantial influence on runtimes. In cases with no heuristic information and constant parameters, however, the benefit of increasing the number of ants from 16 to 32 is minimal in well-performing conditions (i.e. in cases with $\alpha = 1$ and $\rho \in \{.8, .95\}$). Again, this is not the case when sub-optimal heuristics are provided, but, irrespective of heuristic information, the best RE was achieved for $K = 16$. This indicates that the trade-off between runtimes and solution quality is better served with relatively small colonies. Scheduling K to increase during the early phases of the search proved most promising in the evaluation with parameter schedules. Results regarding solution quality were robust with regard to the time-point of switches in K, indicating some tolerance in the specific scheduling. While these schedules actually lead to longer run-times in situations without heuristic information when combined with the flexible parameter schedule and starting schedules for ρ, they do tend to decrease runtimes when combined with (the preferable) constant settings for ρ during the early stages of the search. Again, the results from conditions with scheduled parameters indicate some leniency regarding the actual iteration at which the number of ants increases, with regards to solution quality. Reducing the number of ants during the late stages of the search substantially decreases runtimes, but comes at the cost of reduced solution quality.

Parameter Scheduling Overall, parameter scheduling did not drastically decrease runtimes without also decreasing solution quality. Especially in situations without heuristic information, the benefit of using parameter schedules is barely detectable. This, coupled with the substantial increase in application complexity, may be enough to recommend using constant parameter settings when not providing heuristic information. In these cases the only promising parameter scheduling concerns increases in K during the beginning of the search to enhance ant communication and decreasing ρ towards the end of the search, both resulting in somewhat shorter runtimes. In cases with heuristic information, this may be different, however. Especially early scheduling of β shows promising results in guiding the search through early the phase before deemphasizing the importance of heuristic information in later stages. In all situations, the rigid parameter schedule shows much more consistent runtimes (as is to be expected), but the two approaches are nearly indiscernible in terms of solution quality. However, results from the scheduled parameters with sub-optimal heuristic information show a potential benefit of using the *fps*: when using an early ρ schedule, the *fps* condition outperforms its *rps* counterpart. This is most likely due to a too early decrease in the evaporation coefficient, something that is a risk in an application in which the search space is not well known.

Recommendations Globally optimal parameter settings for the `stuart` approach are somewhat difficult to derive, given the differences in possible applications due to problem size and other instance specifics (e.g. an abundance of improper solutions). However the following seven recommendations are given for applications:

1. Heuristics should be provided when they are sufficiently good and when the search must be guided during the early phases.

The evaluation shows that the search often results in good solutions even when sub-optimal heuristics are provided. In applications, heuristics are likely to be closer to the optimal condition shown here than to the sub-optimal condition. Recall that the sub-optimal heuristics were derived from the median solution and were not elicited in any reasonable fashion. However, the evaluation also shows that heuristics are not necessary to construct close-to-optimal solutions in a situation in which very few problems occur in early search. Therefore, the use of heuristics is recommended in situations in which their quality is sufficiently high. Beyond that, some instances may necessitate heuristics, due to problems in the early search (most likely due to improper solutions stemming from non-positive-definite model matrices). In these cases, it is recommended to schedule β to decrease during the search process.

2. When using heuristics it is recommended to use the global-best deposit-rule.

Across all conditions of the evaluation, the global-best deposit-rule proved to be much faster. In cases with heuristic information it was not noticeably worse than the *ib* deposit rule (when using other parameter settings that are recommended) and had the additional benefit of being less susceptible to the influence of sub-optimal heuristics.

3. In applications without heuristics the iteration-best deposit-rule is recommended.

In this evaluation, the *ib* deposit-rule resulted in demonstrably better solutions when not providing heuristic information. It should, however, be noted that it requires more than twice the runtime of the *gb* deposit rule. In cases in which runtimes are a central concern - i.e. when each CFA takes much longer to estimate than the simple example used here - using the *gb* deposit-rule should be considered.

4. Scheduling parameters is not generally recommended.

The aforementioned scheduling of β to guide the early search is the exception. Scheduling other parameters often proved to be detrimental. Scheduling increases in K during the early search can have a positive impact on runtimes without much loss in terms of solution quality, as does decreasing ρ during the very late search. However, scheduling parameters requires in-depth knowledge about the problem at hand and the risk of unintentionally reducing the solution quality exists. If runtimes are problematic due to model complexity, it is recommended to use the *gb* deposit-rule to reduce runtimes.

5. The non-linearity coefficient α should rarely be different from 1.

The only exception here should be situations in which an application with $\alpha = 1$ showed that there is a substantial need to more strongly discriminate between a plethora of similarly good solutions. In this evaluation $\alpha > 1$ only performed adequate when heuristics were truly sub-optimal, a situation which should rarely be this extreme in applications.

6. A moderate number of ants per colony is sufficient.

Increases in K have a very poor return-on-investment after a minimal K is exceeded. However, defining a specific number is difficult for two reasons. First, the number of appropriate ants depends on the problem-size. Selecting a 30-item scale from a pool of 400 items may require a larger K, though the specific relationship needs further investigation. Second, the proportion of improper solutions increases the number of necessary ants per colony. In this application just over 50% of random solutions were improper. Increases in this ratio will necessitate larger values of K to establish enough viable solutions, which can guide the search via pheromone deposit.

7. Multiple instances of item selection should be run in every single application.

Because the algorithmic approach is based on weighted random selection, different starts may very well result in different solutions. An easy way of improving the chances of finding an extremely good solution is to run multiple instances. This avoids the peril of stagnation in a sub-optimal area of the search space (or running into a local maximum, as it is more commonly known) by initiating multiple starts.

CHAPTER 4

Applications

This chapter will introduce specific examples of the application of the `stuart` approach to data stemming from different studies, each with its own specificities which need consideration.

In addition to the the applications discussed in this Chapter, the first application of the `stuart` approach is provided in the evaluation performed in Chapter 3. Specifically, Section 3.2 introduced the Ryff-Scale and its theoretical structure in accordance to Ryff (1989), Section 3.3 described the optimization problem provided by the reduction of this scale, and Section 3.4 provided three different solutions, including the optimal solution according to the `stuart` approach. In this application data were cross-sectional, stemmed from one group, and were provided by one source of information. In many parameter constellations, the `stuart` approach was able to construct a very good final solution, providing an adequately fitting model for the Ryff-Scale with its six specific facets.

This chapter will provide details on the item selection in more intricate data constellations, beginning with the item selection for a mood-assessment scale in a longitudinal setting. The second application stems from personality psychology and aims at constructing an extremely short scale for the assessment of the Big Five in a cross-cultural application. The third application will present the item selection for a scale assessing emotional expressivity using self- and peer-ratings. The presentation of these three applications will be in line with that provided in Sections 3.2 to 3.4, whereby a brief introduction into the scale itself will be given, followed by a description of the associated problem representation and a quick overview of possible referential solutions. The quality of each constructed scale will be validated and discussed.

4.1 Item Selection in Longitudinal Studies: Application in Mood Assessment

This section will show how a short-form of the Multidimensional Mood State Questionnaire (MDBF; Steyer, Schwenkmezger, Notz, & Eid, 1997) can be constructed, reducing the original pool of 58 items to a final version consisting of twelve items. The MDBF is a questionnaire used to assess the current mood of participants. Mood is conceptualized as a facet of subjective well-being (Eid & Diener, 2004) and can be differentiated from emotion mainly in that it is more stable and not only a direct reaction to situational factors (Steyer et al., 1997). However, it is less stable than other facets of subjective well-being, such as life satisfaction. As such, mood is often used as an outcome in psychological research aimed at understanding situational influences on well-being. In particular, the MDBF has been used in studies investigating a wide array of factors influencing well-being, such as exam taking (Berger & Freund, 2012), physical activity (Fritz, Halfpaap, Grahl, Kirkland, & Villringer, 2013), and confrontation with negative life events (Korn, Sharot, Walter, Heekeren, & Dolan, 2014).

The data used here stem from the original study to construct the scale and are publicly available (Steyer, Schwenkmezger, Notz, & Eid, 2004). 503 subjects were asked to report to what extent each of the 58 original items described their current mood on a 5-point Likert-Scale with the extremes “1 - not at all” and “5 - very much”. Appendix B.2 provides the full MDBF questionnaire. 292 (58.052%) of participants were female and the mean age was 31.191 years with a range from 17 to 78. The MDBF was assessed at four occasions, approximately three weeks apart. More details on the sample and the recruitment procedure are provided by Steyer et al. (2004).

The original item pool consists of 58 adjectives describing the current mood. These items were selected from a general pool of adjectives used in different previous scales for the assessment of mood by a group of experts. Steyer, Schwenkmezger, Eid, and Notz (1991) provide a detailed description of this procedure. After preliminary analysis, items were selected to represent three bipolar mood dimensions: (*a*) good vs. bad, (*b*) awake vs. tired, and (*c*) calm vs. nervous. These facets are conceived of as correlated but distinguishable dimensions of current mood. The bipolarity of the dimensions is an important aspect of the questionnaire. Each facet in the final scales presented by Steyer et al. (1997) consists of item presenting the positive as well as the negative pole of a dimension.

While the original item selection was based on the data from all four measurement occasions in this data set (Steyer et al., 1997), only the first three occasions will be used during item selection in this application. The fourth will be used for validation, with the aim of investigating whether the qualities of the selection made based on the first three occasions can be replicated at a later occasion.

Table 4.1: Theoretical factor structure of the original MDBF item pool.

Facet	Pole	Item Number												No. of Items
Good	Positive	2	5	9	14	17	26	35	37	46	47	56		11
	Negative	13	20	25	28	29	32	34	36	39	45			10
Awake	Positive	1	6	18	22	42	44	48	55					8
	Negative	7	12	21	31	38	40	41	58					8
Calm	Positive	23	30	43	49	52	54	57						7
	Negative	3	4	10	11	15	16	19	24	27	33	50	51	12

4.1.1 Problem Representation

As pointed out in Chapter 2 and then illustrated for the item selection of the Ryff-Scale in Section 3.3, the main challenge in applying the `stuart` approach is representing the problem as the triple $(\mathcal{S}, f, \Omega)$ in each specific instance. Like in Section 3.3, the constraints Ω are discussed first, because they imply the set of possible solutions in $\mathcal{S}$.

The three general constraints in Ω are given by the IMKAR as:

(ω_1) the sum of weights does not exceed capacity - Equation (1.10),

(ω_2) items are selected specifically in their respective facets - Equation (1.12), and

(ω_3) items may be assigned to only one facet in a solution - Equation (1.13).

The first constraint requires two additional parameters: the weight of the items and the capacity of the facets. For this case, the weights of items are all assumed to be equal as $w_{im} = 1$. In accordance to the two short forms of the MDBF presented by Steyer et al. (1997), the constructed scale should consist of twelve items, four for each of the three dimensions. Note that the number of items of the original full MDBF is 24. However, the original scale consists of two alternate short versions each consisting of twelve items. Therefore, $a_m = 4$ gives the final piece required to impose constraint ω_1.

Constraints ω_2 and ω_3 comply with the original selection of adjectives for each of the six general terms (two poles per dimension) as described by Steyer et al. (1991). Table 4.1 provides an overview of the items and their dimensional allocation. Appendix B.2 shows the full 58 item questionnaire. Only 56 of the total 58 adjectives are allocated to the scales, the remaining two (item 8 “ärgerlich [angry]” and item 53 “ängstlich [afraid]”) are not considered for the final scale.

Due to the longitudinal structure of the data, a fourth constraint is imposed as: (ω_4) the same items are selected for repeated measures of the same construct - as given by Equation (2.32). As shown in Sections 2.3.3 and 2.3.5, this implies that there are $M = 9$ facets (three dimensions, each measured three times) assigned to $V = 3$ sets of repeated measures $\mathcal{R}_v$, which are equal to $H = 3$ unique partitions $\mathcal{Q}_h$. Thus, the set $\mathcal{S}$ contains a total number of 4.22×10^{10} possible

Table 4.2: Filter matrix $\boldsymbol{F}_m$ containing the binary heuristics for the MDBF subscale "awake".

	1	6	7	12	18	21	22	31	38	40	41	42	44	48	55	58
1	0	0	1	1	0	1	0	1	1	1	1	0	0	0	0	1
6	0	0	1	1	0	1	0	1	1	1	1	0	0	0	0	1
7	1	1	0	0	1	0	1	0	0	0	0	1	1	1	1	0
12	1	1	0	0	1	0	1	0	0	0	0	1	1	1	1	0
18	0	0	1	1	0	1	0	1	1	1	1	0	0	0	0	1
21	1	1	0	0	1	0	1	0	0	0	0	1	1	1	1	0
22	0	0	1	1	0	1	0	1	1	1	1	0	0	0	0	1
31	1	1	0	0	1	0	1	0	0	0	0	1	1	1	1	0
38	1	1	0	0	1	0	1	0	0	0	0	1	1	1	1	0
40	1	1	0	0	1	0	1	0	0	0	0	1	1	1	1	0
41	1	1	0	0	1	0	1	0	0	0	0	1	1	1	1	0
42	0	0	1	1	0	1	0	1	1	1	1	0	0	0	0	1
44	0	0	1	1	0	1	0	1	1	1	1	0	0	0	0	1
48	0	0	1	1	0	1	0	1	1	1	1	0	0	0	0	1
55	0	0	1	1	0	1	0	1	1	1	1	0	0	0	0	1
58	1	1	0	0	1	0	1	0	0	0	0	1	1	1	1	0

solutions. In line with the evaluation performed by Steyer et al. (1997) for the original item selections, strong factorial invariance is assumed across measurement occasions.

The original scale incorporates the bipolar nature of the MBDF dimensions by retaining the same number of items for each of the poles per dimensions to ensure balance between positive and negative mood in the assessment (Steyer et al., 1997). This balance on the MDBF dimensions can be achieved in two ways in the `stuart` approach. First, the poles can represent different sources of information for the same dimension, making each pole its own facet and determining the consistency between the poles via the MTMM approach discussed in more detail in Section 2.3.4. An application using the MTMM approach in item selection is shown in Section 4.3. In this application an alternative is used by localizing pheromones to arcs instead of nodes as discussed in Section 2.2.2.

To achieve balance between the positive and negative poles of each dimension, heuristics can be used to filter combinations of items by placing binary heuristics on arcs. Table 4.2 shows the binary heuristics for the dimension "awake". The rows represent the starting point of an arc, the columns the end point. Thus, initially selecting item 1 eliminates all items indicating the positive pole of "awake" from the next selection. Using Equation (2.12), the selection probability of an item i' stemming from the same pole as the item i previously selected is given by:

$$p(x_{(i,i')m} = 1|t) = \frac{[\phi_{(i,i')m}(t)]^{\alpha} 0^{\beta}}{\sum\limits_{i'=1}^{I_m} [\phi_{(i,i')m}(t)]^{\alpha} [\eta_{(i,i')m}]^{\beta}} = 0$$

and that of selecting an item i'' from the opposite pole is given by

$$p(x_{(i,i')m} = 1|t) = \frac{[\phi_{(i,i'')m}(t)]^\alpha 1^\beta}{\sum\limits_{i''=1}^{I_m} [\phi_{(i,i'')m}(t)]^\alpha [\eta_{(i,i'')m}]^\beta} = \frac{[\phi_{(i,i'')m}(t)]^\alpha}{\sum\limits_{i''=1}^{I_m} [\phi_{(i,i'')m}(t)]^\alpha}.$$

Thus, even though heuristics are provided in **H**, the actual search process simplifies to a case without heuristics and the parameter settings recommended in Section 3.9 for situations without heuristics should be considered.

The final component needed for the problem representation is the objective function $f(s)$. As was done for the application regarding the Ryff-Scale in Equation (3.3), the objective function is defined as

$$f(s) = \begin{cases} \Phi(s), & \text{if } s \in \mathcal{S}^* \\ 0, & \text{else} \end{cases}, \tag{4.1}$$

where $\mathcal{S}^*$ is the subset of viable solutions, i.e. solutions converging to results with positive-definite latent, residual, and manifest covariance matrices. The pheromone function $\Phi(s)$ is defined as

$$\Phi(s) = \frac{1}{1+e^{-10(M_{rel_{ms}}-0.8)}} + \left(.5 - \frac{.5}{1+e^{-100(\text{RMSEA}_s-.05)}}\right) + \left(.5 - \frac{.5}{1+e^{-100(\text{SRMR}_s-.05)}}\right). \tag{4.2}$$

Here, $M_{rel_{ms}}$ is the average reliability over all facets in a given solution. This is used here instead of composite reliability, because the repeated measures make a composite nonsensical across all facets. As discussed in Section 2.1, this pheromone function implies the highest discrimination between solutions at average facet reliabilities of .8. This value is chosen because Steyer et al. (1997) show that the internal consistencies (computed via Cronbach's α) of each 4-item facet lie between .73 and .89. For RSMEA as well as SRMR, the conservative values of .05 are used for the highest discrimination. In line with the limits derived for the objective function used in the shortening of the Ryff-Scale in Section 3.3, the limits of $f(s)$ can be determined by simple computation. Using the worst-case values ($M_{rel_{ms}} = 0$, RMSEA $= \infty$, SRMR $= 1$) the lower limit is $\min \Phi(s) \approx 0$ for the pheromone function, while $\min f(s) = 0$, due to the definition of $f(s)$ in Equation (4.4). The upper limit, computed with $M_{rel_{ms}} = 1$, RMSEA $= 0$, SRMR $= 0$, is $\max f(s) = \max \Phi(s) = 1.874$.

Thus, with these settings in place, the `stuart` approach will search for a twelve item solution with high composite reliability and close approximate fit, as indicated by the RMSEA and the SRMR, while allowing only combinations consisting of the same amount of positive and negative items (as imposed via $\mathbf{H}_m$). Pheromones are localized to arcs.

Table 4.3: Items for the three dimensions of the MDBF in the s_A^{ts} with their Reliabilities at each of the first three occasions. Items from the negative poles are *emphasized.*

					Reliability Coefficients		
Subscale	Items				Occ. 1	Occ. 2	Occ. 3
Good	2	14	*28*	*32*	0.835	0.882	0.890
Awake	1	*31*	*41*	55	0.853	0.883	0.897
Calm	*3*	*19*	23	57	0.754	0.802	0.830

Table 4.4: Items for the three dimensions of the MDBF in the s_B^{ts} with their Reliabilities at each of the first three occasions. Items from the negative poles are *emphasized.*

					Reliability Coefficients		
Subscale	Items				Occ. 1	Occ. 2	Occ. 3
Good	*29*	*39*	47	56	0.855	0.886	0.867
Awake	*38*	*40*	42	48	0.862	0.884	0.886
Calm	*4*	*10*	30	49	0.723	0.822	0.828

4.1.2 Original MDBF

As was the case in Section 3.4 for the Ryff-Scale, the original selection of items can be viewed as a referential solution against which to contrast the performance of the selection made with the `stuart` approach. However, as stated above, the original MBDF uses 24 items in the full scale, but consists of two halves which can be used separately (Steyer et al., 1997). Thus, these two twelve item versions will both be investigated as referential solutions.

The item allocation of the first original MDBF half (referred to as s_A^{ts}) is provided by Table 4.3. The reported reliabilities are computed in line with Yang and Green (2010). This solution provides a model with unconvincing fit (χ^2 = 2230.244, *df* = 594, $p < .001$, RMSEA = 0.074, SRMR = 0.057, CFI = 0.852) when incorporating three measurement occasions and strong factorial invariance over time. Using the quality function given in Equations (4.4) and (4.5), this solution achieves $f(s_A^{ts}) = 0.821$.

The second original version (denoted s_B^{ts}) is shown in Table 4.4. As was the case for s_A^{ts}, the CFA for s_B^{ts} provides a model with sub-optimal fit (χ^2 = 2339.966, *df* = 594, $p < .001$, RMSEA = 0.076, SRMR = 0.058, CFI = 0.84). Computing the objective function of this solution via Equation (4.4) gives $f(s_B^{ts})$ = 0.806, which is extremely similar to the quality determined for version A of the MDBF, indicating the success of the aim of creating two similar versions of this scale.

To locate the quality of these two solutions in the overall space of possible solutions, 10000 random combinations were generated and evaluated. Note that these solutions were not truly random, because they were required to adhere to the filter matrices $\boldsymbol{F}_m$, which ensure an equal number of positive and negative items for each facet. Of the random solutions, 4133 (41.33%)

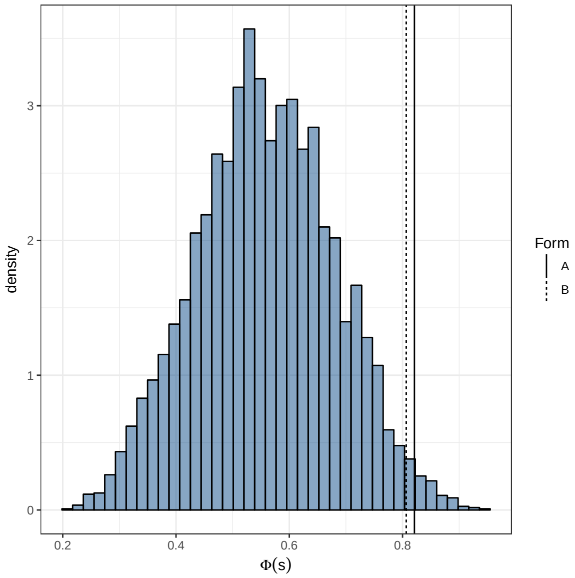


Figure 4.1: Histograms of the pheromone function of the random solutions. The dashed lines represent the quality achieved by the original solutions s_A^{ts} and s_B^{ts}.

resulted in $f(s) = 0$ due to non-viable solutions. The historgram of the $\Phi(s)$ of the remaining solutions is shown in Figure 4.1, with the vertical lines representing the quality of the two MDBF forms. The best random solution resulted in $f(s) = 0.946$, given by a CFA with mediocre fit ($\chi^2 = 1907.909$, *df* $= 594$, $p < .001$, RMSEA $= 0.066$, SRMR $= 0.05$, CFI $= 0.875$). The facet reliabilities ranged from 0.724 to 0.893, thus providing a similar overall solution quality as the two item combinations used in s_A^{ts} and s_B^{ts}.

4.1.3 Item Selection

Using the problem representation discussed in Section 4.1.1, the `stuart` approach was applied to the first three measurement occasions of the MDBF assessment. As mentioned above, only 56 of the 58 items were deemed eligible for selection, because the two remaining items were not attributable to any of the three theoretical dimensions.

In line with the results derived from the evaluation of the `stuart` approach in Section 3.7.1 and the recommendations made in 3.9, a fixed parameter schedule with an iteration-best pheromone deposit was chosen. The iteration-best deposit rule was chosen, because it proved to result in better solution quality, while being much slower than the *gb* deposit rule. Because the problem is relatively small in this instance, runtimes are not expected to be of substantial interest. In line with the results presented in Section 3.7.1, the non-linearity coefficient was chosen to be $\alpha = 1$. The evaporation coefficient was set to $\rho = .95$ to avoid premature convergence and $K = 32$ ants were used throughout the entire search. As described above, pheromones were set to be localized

Table 4.5: Items for the three dimensions of the MDBF in the s^{gb} with their Reliabilities at each of the first three occasions. Items from the negative poles are *emphasized.*

					Reliability Coefficients		
Subscale	Items				Occ. 1	Occ. 2	Occ. 3
Good	2	14	*28*	*34*	0.848	0.890	0.900
Awake	1	*38*	*41*	42	0.861	0.882	0.878
Calm	*10*	*19*	49	57	0.776	0.821	0.840

to arcs, increasing the search complexity considerably. Because of this, larger values for K and ρ were chosen than indicated by the evaluation results for the item selection in the Ryff-Scale. Note that pheromones are initialized to ϕ^{max} irrespective of their heuristic information, meaning that pheromones on non eligible choices must also evaporate - requiring the minimum runtime given by Equation (3.2). Appendix C.1 provides the annotated R-Syntax for the item selection performed here.

The search using these parameters required 601 iterations (totaling 19232 CFA estimations) and took 170.865 minutes to complete utilizing R version 3.3.2 (R Core Team, 2016) and lavaan version 0.5-22 (Rosseel, 2012) on a machine with an Intel Core i7-5600U Quadcore CPU running Ubuntu 16.04.

The items selected in s^{gb} are shown in Table 4.5. Of these twelve items, eleven are also used in one of the two original short forms, with the only exception being item 34 (“in gedrückter Stimmung [in low spirits]”), an indicator of the good vs. bad facet. The model incorporating strong factorial invariance for three measurement occasions showed sub-optimal model fit ($\chi^2 =$ 1665.138, $df = 594$, $p < .001$, RMSEA = 0.06, SRMR = 0.048, CFI = 0.9). In combination with the facet reliabilities shown in Table 4.5 this amounts to $f(s^{gb}) = 1.047$.

It should be noted that despite the model fit of the s^{gb} not being optimal, both RMSEA and SRMR fall into the range of acceptable model fit (Hu & Bentler, 1999; Brown, 2015). As discussed by Moshagen (2012), model fit, as assessed via indicators that are based on the the fit-function, depends substantially on the size of the covariance matrix, i.e. the number of manifest variables included in a model. In this case there are 36 Items used to indicate a total of 9 latent variables, which may drastically reduce the values of the RMSEA and CFI. The SRMR is not directly affected by this, because it is not based on the fit function, but instead on residuals. Marsh, Hau, and Wen (2004) as well as Brown (2015) propose using model comparisons and in-depth analysis of specific misfit, in addition to overall fit indexes, to determine whether a model is suitable in a given situation. In this instance modification indices are investigated and four alternative models are estimated to determine possible reasons for and the degree of model misfit of the final solution, to determine whether the proposed measurement structure used here for the MDBF is suitable.

The modification indices identify restrictions on the residual correlations as the main sources

Table 4.6: Fit criteria of the five alternative models for the selected MDBF items. The five models are denoted "Original" for the model used in the item selection, "Weak Inv." for weak measurement invariance, "Conf. Inv." for configural invariance, "Auto Cor." for a model allowing for all auto-correlations of items, and "CTC(M-1)" for a model incorporating the CTC(M-1) approach.

	Original	Weak Inv.	Conf. Inv.	Auto Cor.	CTC(M-1)
N. Par.	108	126	144	144	225
χ^2	1665.138	1623.254	1596.654	1323.196	747.206
df	594	576	558	558	477
p	0.000	0.000	0.000	0.000	0.000
RMSEA	0.060	0.060	0.061	0.052	0.034
SRMR	0.048	0.047	0.045	0.047	0.030
CFI	0.900	0.902	0.903	0.929	0.975
AIC	46483.824	46477.940	46487.340	46213.882	45799.891
BIC	46939.647	47009.735	47095.105	46821.647	46749.524
aBIC	46596.846	46609.800	46638.037	46364.579	46035.355

of misfit. Of these, suppressed correlations between items with the same valence (i.e. positive or negative adjectives) at the same measurement occasion show the most potential for improving overall model fit. To illustrate, the total sum of modification indices in the s^{gb} is 3377.578, of this 1387.034 (41.066%) are due to restrictions placed on residual correlations at the same measurement occasion. In contrast, 393.075 (11.638%) are due to restrictions placed on the autocorrelation of items. The sum of modification indices accrued by restrictions due to measurement invariance was almost negligible. Note, that these numbers are purely illustrative.

A more direct approach is given if the four less restrictive models are estimated and compared to the original model, to show which general conceptual modification to the measurement model might be necessary. Table 4.6 shows a selection of fit criteria for the five models. The model allowing for the auto-correlations of items imposed strong measurement invariance. The CTC(M-1) approach is realized by setting the positive poles as the reference method and defining additional, dimension specific residual factors for the items assessing the negative poles. Note that in the original presentation of the MDBF scale by Steyer et al. (1997), EFA revealed a fourth, general factor. This fourth factor is interpreted as representing a general response style (with higher values indicating more extreme responses), but not included in this model due to the substantial challenges associated with bi-factor CFA models (e.g. Eid, Geiser, Koch, & Heene, 2016). Instead, the CTC(M-1) structure is used here, as a model accounting for the lack of unidimensionality in the assessment.

While a direct model comparison via the Likelihood-Ratio-Test (LRT) shows the restrictions leading from weak to strong invariance to lead to significantly more misfit ($\chi^2 = 41.883$, $df = 18$, $p = 0.001$), the overall fit does not decrease notably by incorporating these assumptions. In fact, both the BIC and its sample-size adjusted version shown in Table 4.6 prefer the model incorporating strong invariance, indicating how similar the fit of these two models is. Beyond that, the restrictions associated with the weak invariance assumption does not significantly worsen model fit when compared to the configural invariance model ($\chi^2 = 26.6$, $df = 18$, $p = 0.087$).

A somewhat different picture emerges when comparing the original model to the one allowing for the auto-correlation of the residuals. A direct comparison via the LRT suggests that suppressing these correlations may be too strict an assumption for the measurement model of the MDBF ($\chi^2 = 341.941$, $df = 36$, $p < .001$). The main implication of a model allowing for residual correlations is a weakening of the unidimensionality assumption with regards to the time-stable components. If items are more strongly correlated with themselves than is suggested by incorporating only the correlations of the latent variables over time, this may imply different stability for certain subdimensions within the dimensions that are defined at each single occasion of measurement. On the other hand, this may also be due to stable measurement bias introduced by, for example, item phrasing. In this case it is reasonable to suspect that this bias is introduced by the enforced bipolarity of the items assessing these three dimensions. These residual correlations lie in the range of $[0.028; 0.259]$ and their average is 0.151.

As shown in Table 4.6, the most drastic improvement in model fit is achieved if the CTC(M-1) is introduced to accommodate for the bipolarity in the assessment of the dimensions. This is not surprising given the increase in model parameters (the model requires an additional 117 parameters when not imposing additional assumptions). However, this model results in a non-positive definite latent covariance matrix due to auto-correlations of method factors, depicting the negative pole of the "awake" dimension, exceeding 1 and thus, does not constitute a real alternative. However, to inform a possible change in the conceptual measurement model it can be informative to investigate the degree to which items are specifically measures of the negative pole - i.e. their method specificity (c.f. Eid, Lischetzke, Nussbeck, & Trierweiler, 2003). In this case the method specifities range from 0.11 to 0.507, with the highest observable for the "calm" dimension. This, in combination with relatively high stability of the negative-pole specific effects, indicates the possibility of rethinking the measurement model of the MDBF to incorporate these phrasing specific effects.

Despite these possible extensions and alterations of the basic MDBF measurement model, it should be noted that the original model showed adequate fit. The extension to a CTC(M-1) structure to include the specific factors for the negative poles may be viewed as excessive, because of the immense increase in model complexity. Additionally, simple adjustments in terms of measurement invariance do not improve the model drastically. Thus, the basic model is kept and validation is attempted for the fourth measurement occasion.

4.1.4 Selection Validation

To determine whether the item selection is successful in producing a scale with results that are replicable, the fourth occasion of the original data collection was not used in the selection process. Instead, it is used here as a longitudinal version of a validation sample. Three steps are used to appraise the quality of the solution s^{gb}: (a) results from the three occasion specific models are compared to those of the holdout occasion, (b) the fourth occasion is included in a longitudinal model to determine whether the invariance holds for the fourth occasion, and (c) the selection is compared to an occasion specific selection performed at the fourth occasion.

Figures 4.2 and 4.3 show the path diagrams of the measurement models for all four occasions with the unstandardized estimates. A glance at the factor loadings reveals very similar patterns for the last three occasions, while the first occasion seems somewhat different. This is in line with to so-called "socratic effect" (McGuire, 1960; Rosen & Wyer, 1972; Wyer, 1974), which implies that measures taken at later occasions in a study are more homogeneous, due to the participants' desire to achieve cognitive consistency with regards to the constructs that are being assessed. This may indicate that a questionnaire, such as the MDBF, changes its measurement structure slightly with participants' increasing familiarity with it, becoming more reliable and unidimensional at later occasions in a longitudinal study. This could be interpreted to mean that the first measurement occasion should not be included in the process of item selection, if the goal of scale construction is a measure which is ideal for longitudinal assessments. However, the argument for its inclusion is that, in most studies, the first measurement occasion is vitally important and not merely to acquaint participants with the assessment material. Therefore, including the first occasion in the process of item selection will result in a final questionnaire that contains those items that are adequate for both, longitudinal invariance and inclusion of the first occasion, though likely not the optimum for either.

Table 4.7 shows the model fit of the occasion specific models. Astonishingly, the measurement model does not achieve tolerable fit at any of the occasions, despite the overall model for the three occasions investigated in the previous section being adequate. It should be noted that, despite this degree of misfit, the selection performed via the `stuart` approach outperforms both forms presented by Steyer et al. (1997).

Including the fourth occasion into the longitudinal model results in a model with adequate approximate global fit (RMSEA = 0.058, SRMR = 0.048), though the comparison with the independence model does not indicate a suitable measurement model (CFI = 0.877). In terms of exact fit, the model is rejected by the χ^2-Test ($\chi^2 = 2875.385$, $df = 1068$, $p < .001$). When assuming strong factorial invariance for the first three occasions, model comparisons testing for weak vs. configural invariance ($\chi^2 = 16.039$, $df = 9$, $p = 0.066$) as well as strong vs. weak invariance ($\chi^2 = 20.882$, $df = 12$, $p = 0.052$) of the fourth occasion did not lead to rejection of the invariance assumptions. Although the direct comparison between configural and strong

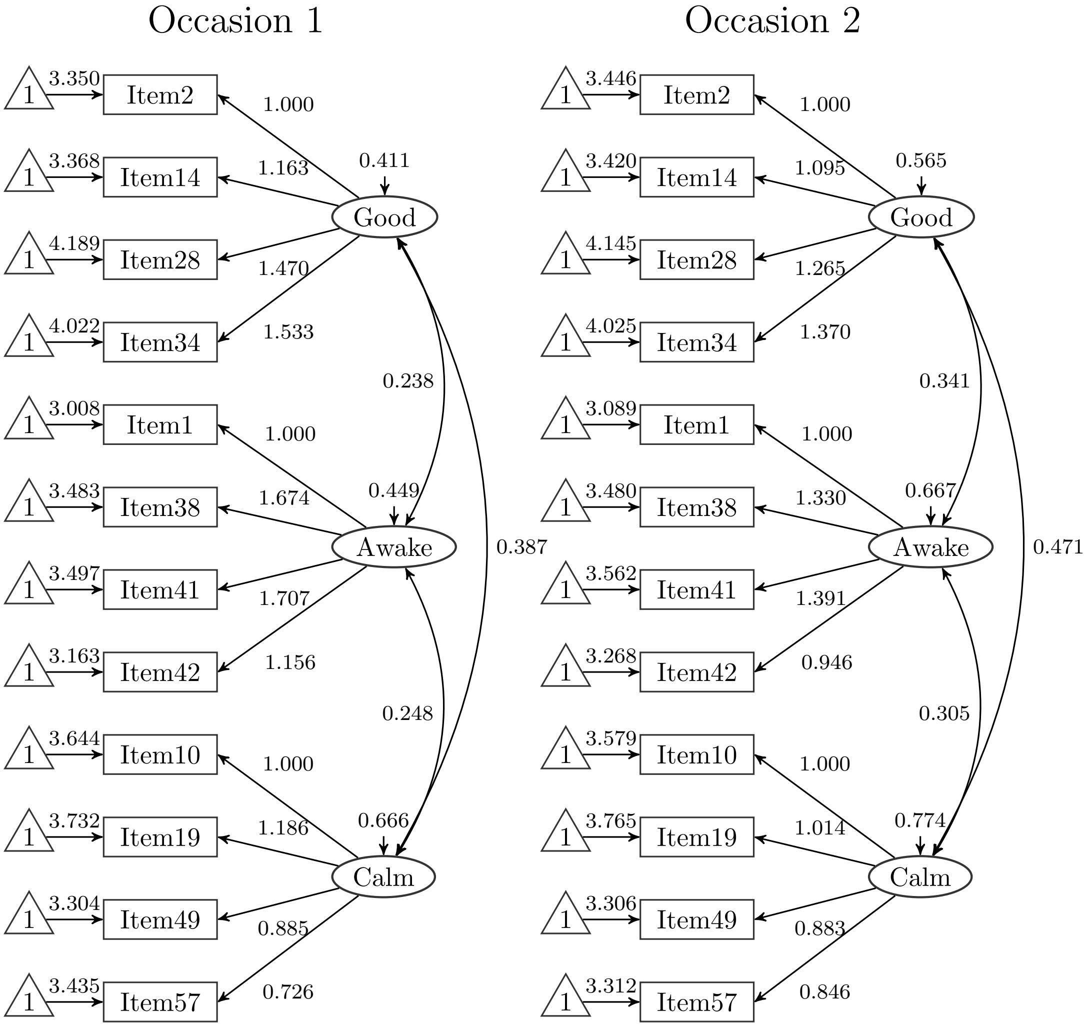


Figure 4.2: Path diagrams of the occasion measurement models for the selected MDBF items. Residuals are not depicted, because they were not constrained in the longitudinal model.

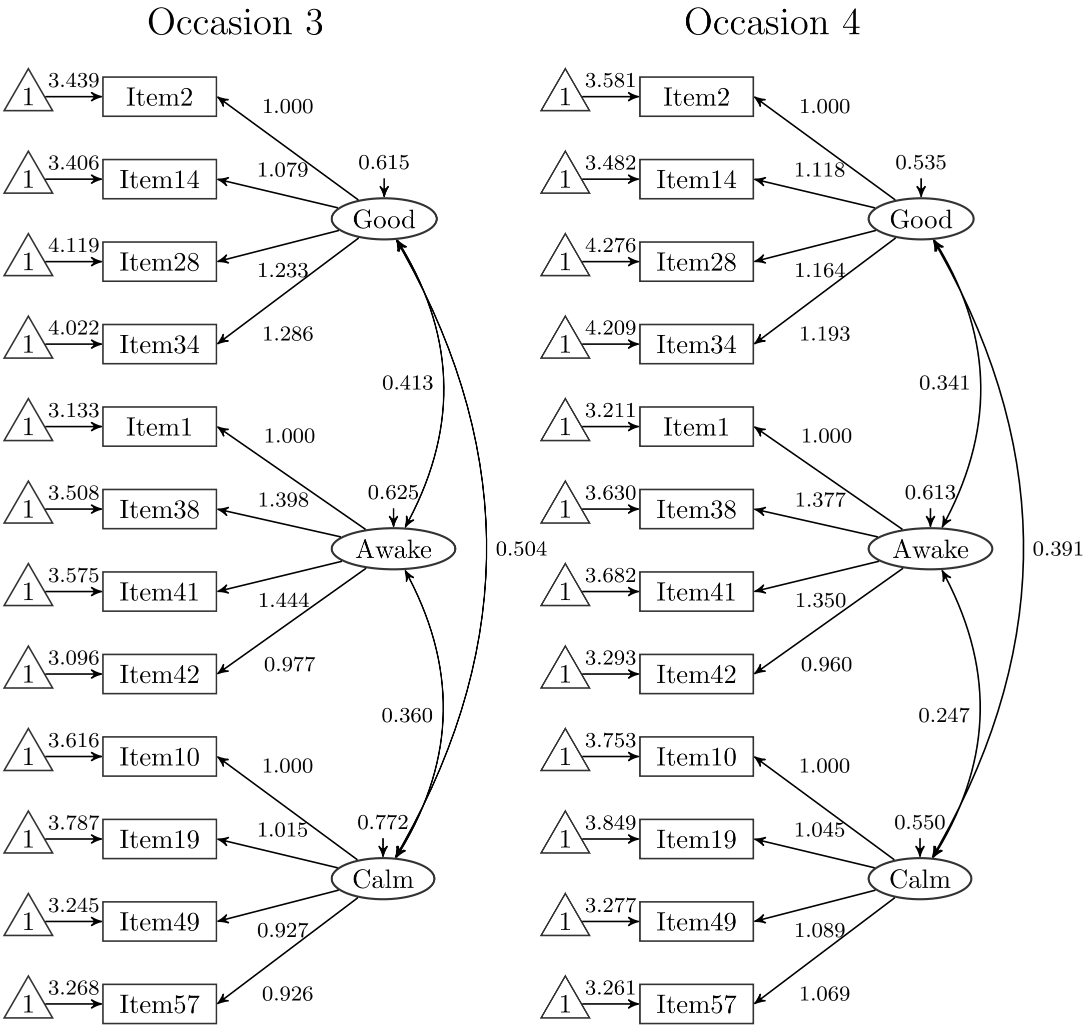


Figure 4.3: Path diagrams of the occasion measurement models for the selected MDBF items. Residuals are not depicted, because they were not constrained in the longitudinal model.

Table 4.7: Fit criteria of the four occasion specific models for each set of selected MDBF items.

		Occ. 1	Occ. 2	Occ. 3	Occ. 4
s^{gb}	χ^2	277.611	265.543	284.852	363.641
	df	51	51	51	51
	p	0.000	0.000	0.000	0.000
	RMSEA	0.094	0.091	0.095	0.110
	SRMR	0.053	0.040	0.042	0.053
	CFI	0.920	0.938	0.938	0.906
	AIC	16425.791	15551.476	15285.014	15334.280
	BIC	16261.188	15386.873	15120.411	15169.677
	aBIC	16302.001	15427.686	15161.225	15210.491
s^{ts}_A	χ^2	426.053	414.504	517.171	482.462
	df	51	51	51	51
	p	0.000	0.000	0.000	0.000
	RMSEA	0.121	0.119	0.135	0.130
	SRMR	0.071	0.062	0.069	0.070
	CFI	0.866	0.895	0.886	0.871
	AIC	16452.266	15619.039	15115.485	15411.945
	BIC	16287.663	15454.436	14950.882	15247.342
	aBIC	16328.477	15495.250	14991.696	15288.156
s^{ts}_B	χ^2	480.541	437.064	432.766	460.037
	df	51	51	51	51
	p	0.000	0.000	0.000	0.000
	RMSEA	0.129	0.123	0.122	0.126
	SRMR	0.078	0.069	0.059	0.063
	CFI	0.851	0.893	0.894	0.881
	AIC	16618.443	15555.356	15609.882	15435.541
	BIC	16453.840	15390.753	15445.279	15270.938
	aBIC	16494.654	15431.566	15486.093	15311.752

measurement invariance for the fourth occasion via LRT results in rejection of the invariance assumptions ($\chi^2 = 36.921$, *df* $= 21$, $p = 0.017$), the information criteria unanimously favored the more restrictive solution assuming the equality of factor loadings and intercepts across all four occasions (AIC: 61422.824 vs. 61427.902, BIC: 62081.236 vs. 62174.947, aBIC: 61586.078 vs. 61613.133). As was the case for the original model with three occasions, modification indices suggest that most of the misfit is due to the restrictions placed on residual covariances within each measurement occasion.

Finally, to locate the quality of the solution in the space of possible solutions, 10000 pseudo-random solutions were constructed using only the fourth occasion. As was the case in the previous section, these solutions are all subject to the constraint that each facet must consist of two positive and two negative items. Of these 10000, a total of 9462 resulted in viable solutions. Additionally, `stuart` was used to search for a selection of items at the fourth occasion using the same parameter settings that were used in the construction presented in Section 4.1.3. This solution showed very similar model fit as the one derived from the first three occasions ($\chi^2 = 281.584$, *df* $= 51$, $p < .001$, RMSEA = 0.095, SRMR = 0.043, CFI = 0.93). The reliabilities of the facets differed by no more than .011 (for the subscale "good"). Figure 4.4 shows the histogram of $\Phi(s)$ of the 9462 viable pseudo-random solutions, with the quality of both solutions included. In terms of the quality defined in Equation 4.4, the occasion specific solution proved slightly better - with $f(s) = 0.956$ than the application of the solution found for the first three occasions, for which $f(s) = 0.848$.

4.1.5 Discussion

This section presented an application of the `stuart` approach to a longitudinal data setting with a filtering approach to heuristic information. Although the final scale may not deliver satisfactory model fit, it does outperform the manual selections with regards to fit and scale reliability. Additionally, it is much easier to implement than the item selection performed by Steyer et al. (1997). Appendix C.1 shows the entire R-code necessary for the application of the `stuart` approach to the MDBF.

As pointed out throughout this section, one of the main challenges with item selection in this instance is the requirement of using the same number of positive and negative items. This leads to subscales that are not truly unidimensional, which, in turn, reduces the fit of the measurement model. Thus, it seems promising to either abandon the idea of unidimensional scales which assess the positive and negative in a balanced fashion, or to acknowledge the differences in the poles by introducing an MTMM structure to the measurement model. An application of the `stuart` approach in a situation with MTMM data structure is shown in Section 4.3. Completely removing heuristic information from the search results in a selection of items that provides a well-fitting model across three measurement occasions ($\chi^2 = 992.373$, *df* $= 594$, $p < .001$, RMSEA =

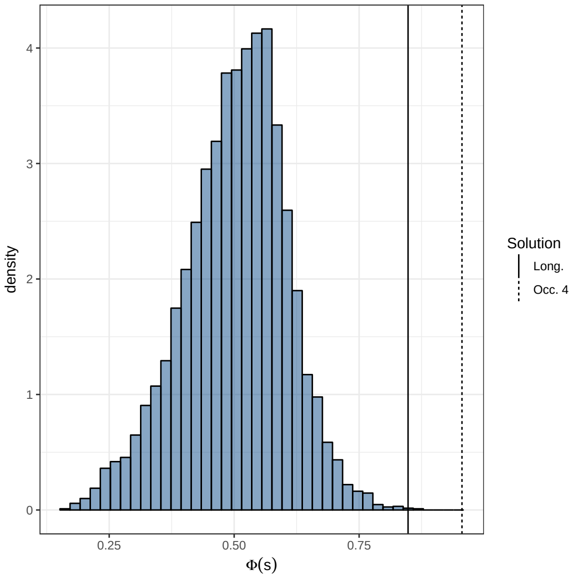


Figure 4.4: Histograms of the pheromone function of the random solutions. The solid line represents the quality achieved by the solution found in the first three occasions, the dashed line the solutions found specifically for the fourth occasion.

0.037, SRMR = 0.037, CFI = 0.971). Unsurprisingly, this scale consists predominantly of items assessing the negative pole of the dimensions of the MDBF (with item 49 “ruhig [calm]” being to sole exception).

4.2 Item Selection in Multiple Groups: Application in Big Five Assessment

This section presents the item selection from the international personality item pool (IPIP; *International Personality Item Pool: A Scientific Collaboratory for the Development of Advanced Measures of Personality Traits and Other Individual Differences*, 2016; Goldberg et al., 2006) to obtain a ten-item scale for the quick and cursory assessment of the Big-Five personality dimensions.

The IPIP is a collection of over 3000 items for the assessment of different personality dimensions. The underlying aim of the project is to “provide rapid access to measures of individual differences, all in the public domain, to be developed conjointly among scientists worldwide” (*International Personality Item Pool: A Scientific Collaboratory for the Development of Advanced Measures of Personality Traits and Other Individual Differences*, 2016), which is why it contains such a wide collection of items related to the assessment of personality. Being in the public domain, the IPIP has spawned over 600 scientific papers and 250 different scales in over

40 different languages. Its origins lie in the works of Hendriks, Hofstee, and Raad (1999), who constructed the Five-Factor Personality Inventory - a scale of 100 items aimed at assessing five general personality factors. This inventory, and the approach of the IPIP in general, is based on the lexical approach and is focused on items describing specific behaviors, rather than general motives and tendencies. Currently, perhaps the most prominent scales based on IPIP components are the 50-item scale proposed by Goldberg (1992) and the 120-item IPIP NEO-PI-R proposed by Johnson (2014).

However, these two, along with many other scales derived from or contributing to the IPIP, have the important characteristic of being extremely long, detailed questionnaires. Johnson's (2014) IPIP NEO-PI-R measures 30 specific subscales of the Big Five with 4 items each, which not just entails a lengthy assessment for study participants, but also complex analysis for the researcher. As a point of general interest in psychology, the assessment of personality traits is often part of studies which focus primarily on other constructs, thus necessitating a short scale providing rough approximations of the Big Five personality traits.

There has been a considerable amount of research on the construction of extremely short scales for the assessment of the Big Five. Gosling, Rentfrow, and Swann (2003) constructed a five- and a ten-item scale, derived mainly from the items constructed by Goldberg (1992), using an exclusively theory guided approach to item selection. Due to the limitations in modeling possibility and the psychometric shortcomings of assessing each trait with only one item, they recommend using the ten-item version, called TIPI (Ten Item Personality Inventory). The TIPI has been shown to lead to adequate results in a CFA framework (e.g. Ehrhart et al., 2009). Similarly, Rammstedt and John (2007) constructed a ten-item version of the BFI-44 (John, Donahue, & Kentle, 1991), also stressing the need for these extremely short assessments of personality traits, especially in online assessments. Another widespread short-version is the Mini-IPIP presented by Donnellan, Oswald, Baird, and Lucas (2006).

In addition to being short, a point of relevance in constructing a scale for use in modern study settings is its cross-cultural applicability. The TIPI, for example, is available in 20 languages and the original article has been cited 1233 times[1], indicating its usefulness to researchers from a wide array of countries and areas of research. As is the case with IPIP, the TIPI is publicly available.

In this application, a ten-item short scale is constructed from a pool of 300 IPIP items assessing a total of 30 different facets of personality - each of the Big Five personality traits Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism is divided into six specific sub-dimensions (Johnson, 2014). To ensure that the ten item short version constructed here has a minimal degree of globality in the dimensions it assesses, item selection is limited in such a way that it ensures that selected items do not assess the same sub-dimension of a personality trait. It is easily possible to select two items from the pool to represent a dimension

[1] As determined via Web of Science, January 10, 2017.

which are, in effect, almost identical. For example items 176 ("Remain clam under pressure") and 296 ("Am calm even in tense situations") are both indicators of neuroticism. However, they are also both indicators of the same subscale, namely "vulnerability". Selecting these two items to assess the global neuroticism function may result in very high reliability (because these two items are bound to be highly correlated) and good overall model fit, but would be a disservice to construct validity. These limitations are imposed via heuristics and are discussed in more detail in the following section.

Data stem from Johnson's (2014) assessment of 307313 participants for the construction of the 120-item IPIP NEO-PI-R and are publicly available via the Open Science Framework at `https://osf.io/tbmh5/`. Participants were self-selected, not actively recruited, and anonymously filled out the questionnaire online. For more details on the assessment procedure see (Johnson, 2014). Due to the online nature of the assessment, participants stem from a wide variety of different countries. Specifically, 238 different countries provided at least one participant - an astonishing feat, given that the number of sovereign states as of 2000 (the year of data collection) was no larger than 213.

To ensure the intercultural applicability of the selected ten-item version constructed in this section, three countries were selected from the pool of countries: Mexico, France, and Malaysia. These three were selected specifically, because they provide roughly the same number of participants ($N_{\text{MX}} = 700$, $N_{\text{FR}} = 854$, $N_{\text{MY}} = 911$)[2] and are in vastly different regions of the earth. A selection was made to ensure empirical identification of the CFA in each country (i.e. an adequate sample size within each group) and to reduce estimation time of the CFAs for this application. Beyond these three countries, South Africa was selected as a fourth country to provide a validation sample. South Africa was chosen specifically, because it provides a large sample ($N_{\text{ZA}} = 927$) and is vastly different from the three countries used to select the ten items.

4.2.1 Problem Representation

In line with Section 1.3, the problem of selecting items from the 300-item pool is represented by the triple $(\mathcal{S}, f, \Omega)$, consisting of the set of possible solutions $\mathcal{S}$, the objective function f, and the constraints Ω. Because the set of possible solutions is subject to the constraints that are imposed, Ω is discussed first.

As shown in Section 1.3.2, the basic constraints associated with all applications of the `stuart` approach are:

(ω_1) the sum of weights does not exceed capacity - Equation (1.10),

(ω_2) items are selected specifically in their respective facets - Equation (1.12), and

[2] The countries' names are abbreviated according to ISO 3166 alpha-2, available at `http://www.iso.org/iso/country_codes`

(ω_3) items may be selected to only one facet - Equation (1.13).

In this application the weights of all items are set to be $w_{im} = 1$ and the capacity of each of the $M = 5$ facets is set to $a_{im} = 2$. This combination ensures the selection of two items per dimension, thus constructing a ten-item scale. Beyond these constraints, the multiple group aspect of the application is handled via:

$$\mathcal{C}_g = \mathcal{C}_{g'} \tag{2.22, repeated}$$

and

$$(\mathcal{S}_g, f_g, \Omega_g) = (\mathcal{S}_{g'}, f_{g'}, \Omega_{g'}), \tag{2.23, repeated}$$

stating the assumption that the itempool consists of the same items for all groups and that the problem is identical for all groups. The latter implies the same constraints, objective function, and set of possible solutions across the three countries. Thus, the item selection is performed independently of the group, and the multiple groups aspect is integrated only in the CFA estimation (as discussed in Section 2.3.2).

Within the CFAs, two invariance assumptions are imposed. First, strong factorial invariance is assumed across nations, thus implying the equality of factor loadings λ_{img} and intercepts α_{img} for all values of g. This assumption allows for the estimation of latent means κ_{mg} in two of the three groups as contrasts to the reference group. In this instance Mexico is chosen as the reference group, but model fit and reliabilities of the final scale are independent of this choice. Second, essentially τ-equivalent measures are assumed within each facet, implying $\lambda_{img} = 1$ for all indicators i, all facets m, and all groups g. This is done to ensure the independence of model identification from the empirical values of the latent correlations, because the different personality facets are not necessarily sufficiently correlated in this collection of items (e.g. Maples, Guan, Carter, & Miller, 2014). In cases with only two indicators it is necessary for latent variables to be correlated to identify a τ-congeneric measurement model without further restrictions - this is not the case if the indicators are constrained to be essentially τ-equivalent (c.f. Brown, 2015).

Because of the size of the item pool, a complete tabulation of the item allocation in the original is not presented here.[3] Each of the Big Five dimensions is assessed with 60 items. As pointed out above, the items assesses not only the five top-level dimensions, but also six specific sub-facets per dimension (Johnson, 2014). Each of these sub-facets is associated with 10 items in the original item pool. Selecting two items for each dimension to construct a 10 item short scale results in a total number of 1.74×10^{16} possible solutions in $\mathcal{S}$.

To ensure that selected items assess different sub-facets, heuristics can be defined as dimension-

[3] A complete item allocation can be found at `http://ipip.ori.org/newNEOFacetsKey.htm`.

specific filter matrices. Because the restriction imposed concerns combinations of items, heuristics (and thereby the item selection in general) must be localized to the arcs between items. This approach was described in Section 2.2.2 and shown for the application of the MDBF in Section 4.1.

In addition to the filter matrices $\boldsymbol{F}_m$, a promising approach to guiding the search in item selection is providing correlations between the items as heuristic information. In this instance, the scales consist of two items and are assumed to be essentially τ-equivalent. Thus, scale reliabilities are closely tied to bivariate correlations and providing these correlations as information in each selection step should guide the search in the direction of more reliable scales. To emphasize stronger correlations and to equally favor strong positive and strong negative correlations, the absolute values of z-Transformed Pearson correlations (Fisher, 1924) are used (denoted $\boldsymbol{Z}_m$) in addition to the filter matrices $\boldsymbol{F}_m$ to obtain the heuristics via

$$\mathbf{H}_m = \boldsymbol{Z}_m \boldsymbol{F}_m. \tag{4.3}$$

In this case the correlations from the entire sample encompassing participants from Mexico, France, and Malaysia are used.

The objective function f is defined in this application as

$$f(s) = \begin{cases} \Phi(s), & \text{if } s \in \mathcal{S}^* \\ 0, & \text{else} \end{cases}, \tag{4.4}$$

with $\mathcal{S}^*$ being the subset of viable solutions. The pheromone function $\Phi(s)$ is defined as

$$\Phi(s) = \frac{1}{1+e^{-25(\min rel_{ms}-0.4)}} + \left(1 - \frac{1}{1+e^{-100(\mathrm{RMSEA}_s-.05)}}\right) + \left(1 - \frac{1}{1+e^{-100(\mathrm{SRMR}_s-.05)}}\right). \tag{4.5}$$

In this instance, the minimal reliability of the five scales is relevant to the quality function instead of the average reliability or the composite reliability, because of the danger of constructing an extremely unreliable dimension with just two items and the remaining four reliabilities quantifying the solution as good, regardless. A reliability of .4 was chosen as the point of strongest discrimination because this is in line with the lowest reliabilities found for the subscales of the TIPI (Romero, Villar, Gómez-Fraguela, & López-Romero, 2012). The pheromone function is limited from below by $\Phi(s) \approx 0$ (in a case with $\min rel_{ms} = 0$, RMSEA $= \infty$, and SRMR $=$ 1) and the objective function is limited by $f(s) = 0$, due to the value assigned to $s \notin \mathcal{S}^*$. The upper limit for both is $\max f(s) = \max \Phi(s) = 2.987$ (for a solution with RMSEA $= 0$, SRMR $= 0$, and $\min rel_{ms} = 1$).

4.2.2 Item Selection

With the problem representation in place, parameters for the search algorithm must be defined. This application provides the search with heuristic information and thus the parameters found to be most promising in Sections 3.7.3 and 3.8.4 are used to guide parameter selection, in this instance.

Both non-linearity coefficients (α and β) are set to 1, because it was the most promising setting for conditions with heuristic information in terms of solution quality. In line with the results presented in Section 3.7.3, the evaporation coefficient is set to $\rho = .95$. Due to the problem size being much larger in this case (even more so due to the localization to arcs) than in the parameter evaluation presented in Chapter 3, the number of ants per iteration is set to $K = 64$ and the maximum number of iterations is set to $T = 512$ after the last s^{gb} was found. Additionally, the *gb* deposit rule is chosen because it proved much faster in Section 3.7.3, with only minute impact on the average and worst-case RD.

Five independent runs of the `stuart` approach were performed, all leading to extremely poor solutions. Closer inspection revealed that 93.482% of all constructed solutions were not viable. The overwhelming percentage of these solutions were not viable because the latent covariance matrix Ψ_g was not positive-definite in at least one of the countries. This was due to facets being indicated by two almost uncorrelated items, thereby leading to negative variance estimates for the latent variables.

A possible solution to this problem is increasing the value of β, thereby making the selection of an item less likely when it is only weakly correlated to the first selected item. This, however, has the drawback of intensifying the differences between acceptable correlations as well, which may lead to too little exploration. Due to the definition of the selection probability defined in Equation (2.12) for cases with localization to arcs, the relative initial selection probability of an item c_2 over an item c_3 after having chosen c_1 is

$$\left(\frac{z_{cor[c_1,c_2]}}{z_{cor[c_1,c_3]}}\right)^{\beta}, \tag{4.6}$$

meaning that, for a value of $\beta = 5$, a path with an associated correlation of .6 would be 56.323 times more likely to be chosen than a path with an associated correlation of .3.

The number of impermissible solutions can be reduced further by imposing another filtering condition. Most improper solutions were due to negative latent variances, caused by a lack of covariance between the two indicators chosen. A simple way reduce the number of these problems is to disallow all combinations of items which are correlated to a sufficiently small extent. In this case $|cor(c_i, c_{i'})| > .2$ is chosen as this cutoff criterion. In the `stuart` approach the coefficient proposed by Yang and Green (2010) is used to determine scale reliability. As reported by Yang and Green (2010, p. 68) this is computed as

$$rel(m_g) = \frac{\sum \Lambda_{mg} \Psi_{mg} \Lambda_{mg}^{\top}}{\sum \Lambda_{mg} \Psi_{mg} \Lambda_{mg}^{\top} + \Theta_{mg}} \tag{4.7}$$

for the facet m in group g. Because of the assumption of essentially τ-equivalent measures, $var(\xi_{mg})$ approximates $cov(c_{img}, c_{i'mg})$ and all $\lambda_{img} = 1$. This means that a two-item facet, for which the correlation between the two items is .2, would approximately achieve a reliability of 0.286. Thus, this additional filter precludes any possible facets with a reliability below this approximate value from being constructed.

In addition to excluding specific combinations of items from the set of possible solutions, the minimum correlation filter reduces the problem-size by removing all items which do not have a correlation of a magnitude larger than .2 with any item not pertaining to the same personality sub-scale. Specifically, this reduces to number of items (from the original size of 60 per Dimension) to 44 for Openness, 59 for Conscientiousness, 54 for Extraversion, 54 for Agreeableness, and 54 for Neuroticism.

In contrast to the item selection for the MDBF presented in Section 4.1, there is no prior, referential solution in this case. Thus, the heuristically favored solution is presented here as one possibility. The heuristically favored solution s^{heu} is most likely to be constructed at the first iteration, when all arcs between items have the same amount of pheromone deposited. In this initial iteration the selection probability, determined via

$$p(x_{(i,i')m} = 1|t) = \frac{[\phi_{(i,i')m}(t)]^{\alpha}[\eta_{(i,i')m}]^{\beta}}{\sum\limits_{i'=1}^{I_m} [\phi_{(i,i')m}(t)]^{\alpha}[\eta_{(i,i')m}]^{\beta}}, \tag{2.12, repeated}$$

is dependent only on the weighted heuristic information provided by Φ_m^{β}. This contains the z-transformed correlations that are $|cor(c_i, c_{i'})| > .2$ between items not stemming from the same sub-dimension. Therefore, the heuristically favored solution is the solution choosing the items that are most highly correlated and from different subdimensions. Table 4.8 shows the items of this solution, with their corresponding dimensions and subdimensions in the original 300-item scale. The correlation (and its Fisher z) pertain to the relation between the two items for each single dimension in the combined sample.

This solution results in a CFA with sub-optimal fit ($\chi^2 = 473.897$, $df = 100$, $p < .001$, RMSEA = 0.067, SRMR = 0.044, CFI = 0.914). Reliabilities are high for subscales consisting of only two items, with an average reliability of 0.654 (Range: [0.494; 0.777]), which is not surprising given the fact that the main selection criterion for item combinations was their correlation.

As pointed out, many of the possible combinations result in non-positive-definite latent covariance matrices in the CFA, making the generation of an informative random sample of solutions much more difficult than was the case for either the Ryff-Scale (Section 3.2) or the MDBF (Section 4.1). In this instance, generating 10000 random samples under the constraints imposed by

Table 4.8: Heuristically favored 10-item scale, selected from the 300-item IPIP. Items were recoded prior to analysis in line with the dimension they assess.

Dimension	Sub-Dimension	No.	Item	Correlation (Fisher z)
Openness	Intellect	143	Enjoy thinking about things.	0.396
	Imagination	153	Spend time reflecting on things.	(0.419)
Conscientiousness	Self-Discipline	145	Carry out my plans.	0.562
	Achievement-Striving	80	Turn plans into actions.	(0.636)
Extraversion	Excitement-Seeking	112	Enjoy being part of a loud crowd.	0.565
	Gregariousness	217	Don't like crowded events.	(0.641)
Agreeableness	Altruism	104	Am concerned about others.	0.422
	Sympathy	149	Am not interested in other people's problems.	(0.451)
Neuroticism	Anger	96	Am often in a bad mood.	0.540
	Depression	131	Have frequent mood swings.	(0.605)

the two filtering conditions described above, results in only 39 (0.39%) viable solutions. Therefore the presentation of random solutions is not deemed appropriate as a benchmark for the quality of the final and heuristically favored solutions.

The final solution constructed under the constraints discussed, is shown in Table 4.9. This solution results in good overall model fit when examining the RMSEA (0.049) and the SRMR (0.039), but the value of the CFI is relatively low (0.915). The latter is most likely due to the low correlations between the items (range: [-0.18, 0.474]), which results in a better fitting fully constrained null-model than in most other cases. The exact test for model fit is, unsurprisingly, significant (χ^2 = 293.631, df = 100, $p < .001$).

The country specific reliabilities of the subscales, as well as the latent means are shown in Table 4.10. The latent means are constrained to zero in the Mexican sample to identify the mean structure of the model. Therefore, the latent means are interpretable as the unidimensional mean differences between the Mexican (the reference group) and the French or the Malaysian sample, respectively. Table 4.10 indicates the significance of these mean differences at $p < .05$.

The reliabilities of the dimensions are somewhat acceptable for dimensions consisting of only two items, ranging from 0.398 (Agreeableness assessed in the French sample) to 0.655 (Neuroticism assessed in the French sample).

Table 4.9: The final 10-item scale, selected from the 300-item IPIP. Items were recoded prior to analysis in line with the dimension they assess.

Dimension	Sub-Dimension	No.	Item
Openness	Intellect	143	Enjoy thinking about things.
	Imagination	153	Spend time reflecting on things.
Conscientiousness	Orderliness	190	Leave a mess in my room.
	Self-Discipline	25	Get chores done right away.
Extraversion	Gregariousness	7	Love large parties.
	Excitement-Seeking	292	Dislike loud music.
Agreeableness	Trust	64	Trust what people say.
	Cooperation	19	Am easy to satisfy.
Neuroticism	Anger	66	Get upset easily.
	Depression	131	Have frequent mood swings.

Table 4.10: Reliabilities and latent means for the dimensions of the IPIP 10-item short scale. Items were recoded prior to analysis in line with the dimension they assess. Means significant at $p < .05$ are *emphasized.*

	Reliabilities			Latent Means		
Dimension	Malaysia	France	Mexico	Malaysia	France	Mexico
Openness	0.493	0.589	0.597	*-0.080*	0.040	0.000
Conscientiousness	0.598	0.545	0.461	*-0.202*	*-0.126*	0.000
Extraversion	0.404	0.510	0.529	*-0.480*	*-0.400*	0.000
Agreeableness	0.446	0.398	0.425	*0.097*	0.014	0.000
Neuroticism	0.632	0.655	0.631	*0.155*	-0.073	0.000

Table 4.11: Fit criteria of the four country specific models.

	Mexico	France	Malaysia	South Africa
χ^2	65.148	72.140	113.626	95.215
df	30	30	30	30
p	0.000	0.000	0.000	0.000
RMSEA	0.041	0.041	0.055	0.048
SRMR	0.035	0.033	0.042	0.034
CFI	0.949	0.948	0.892	0.928
AIC	22115.398	26344.264	27840.382	28316.728
BIC	22274.685	26510.511	28008.891	28485.847
aBIC	22163.553	26399.361	27897.736	28374.690

4.2.3 Selection Validation

As discussed previously, the South African sample (with $N_{\text{ZA}} = 927$) is used as a pseudo-validation sample to investigate the applicability of the 10-item short scale in a different country. As was the case for the application to the MDBF, this validation is performed in three steps: (*a*) results from the three country specific models are compared to those of the South African sample, (*b*) the fourth country is included in the multigroup model to determine whether the invariance holds across these different settings, and (*c*) the selection is compared to an selection performed only within the South African sample.

Table 4.11 shows the fit criteria for each of the country specific measurement models. Closeness of fit indices show good fit of the model within the samples from Mexico, France, and South Africa. Surprisingly, fit is not worst in the validation sample, but instead in the Malaysian sample. Here, RMSEA are SRMR are within the boundaries of acceptable fit, but the CFI is worryingly low. Overall, this may indicate the necessity of constructing a different solution, which might be more suited for the application in Malaysia.

Tables 4.12 through 4.15 show the latent correlations in group specific models for the four countries. While the correlations vary in size rather drastically, the patterns across all four nations seem consistent. Note that the latent correlations are not restricted in the multiple groups model, so their differences have no impact on the misfit of the overall model. However, as a part of investigating the cross-cultural applicability of the short-form, it is necessary to ensure that relationships between the different dimensions are relatively stable across settings and are in line with theoretical expectations. Most prominent is the positive correlation between Openness and Neuroticism across all four samples, which runs contrary to the findings in other short scales (e.g. Donnellan et al., 2006; Gosling et al., 2003). This may indicate a too narrow assessment of Openness (as intellectuality) and/or Neuroticism (as emotional instability).

The means of the ten selected items are shown in Table 4.16. Especially the difference

Table 4.12: Latent correlations and variances (in the diagonal) in Mexico.

	O	C	E	A	N
O	*0.394*				
C	-0.200	*0.523*			
E	0.021	-0.209	*0.715*		
A	0.024	0.159	0.181	*0.394*	
N	0.209	-0.431	0.065	-0.457	*0.831*

Table 4.13: Latent correlations and variances (in the diagonal) in France.

	O	C	E	A	N
O	*0.309*				
C	-0.131	*0.674*			
E	-0.076	-0.059	*0.624*		
A	-0.202	0.072	0.155	*0.319*	
N	0.134	-0.249	-0.177	-0.410	*0.852*

Table 4.14: Latent correlations and variances (in the diagonal) in Malaysia.

	O	C	E	A	N
O	*0.257*				
C	-0.227	*0.719*			
E	-0.044	-0.169	*0.423*		
A	0.068	-0.065	-0.032	*0.343*	
N	0.155	-0.182	0.073	-0.151	*0.757*

Table 4.15: Latent correlations and variances (in the diagonal) in South Africa.

	O	C	E	A	N
O	*0.247*				
C	-0.044	*0.710*			
E	-0.192	-0.117	*0.622*		
A	-0.130	-0.029	0.211	*0.270*	
N	0.034	-0.194	0.095	-0.252	*0.891*

Table 4.16: Country specific item means and their differences for the ten selected items. Items were recoded prior to analysis in line with the dimension they assess.

		Mexico	France	Malaysia	South Africa
	Item 143	4.43	4.40	4.33	4.47
Openness	Item 153	4.02	4.21	4.01	4.21
		-0.41	-0.19	-0.32	-0.27
	Item 190	3.01	2.98	2.97	2.87
Conscientiousness	Item 25	3.13	2.95	2.84	2.81
		0.12	-0.03	-0.13	-0.05
	Item 7	3.10	2.74	2.67	2.89
Extraversion	Item 292	3.70	3.27	3.18	3.76
		0.60	0.54	0.51	0.87
	Item 64	3.23	3.19	3.24	3.12
Agreeableness	Item 19	2.95	3.04	3.17	3.17
		-0.28	-0.16	-0.07	0.05
	Item 66	2.90	2.86	3.07	3.03
Neuroticism	Item 131	3.02	2.90	3.16	2.96
		0.12	0.04	0.09	-0.07

between the means of two items pertaining to the same dimension is of relevance to the strong measurement invariance assumption, which is why they are included in the table. As pointed out in Section 2.3.2, strong measurement invariance implies that the mean difference between groups in manifest variables is due to latent mean differences, and thus unidimensional. Large discrepancies in mean differences between items pertaining to the same dimension may indicate country specificity in item difficulty beyond simple latent mean differences. With regards to Extraversion, Agreeableness, and Neuroticism, Table 4.16 indicates some relations which may call the group-independence of the short-scale into question.

A more definite decision regarding the inter-cultural replicability of the factor-structure is given by including the South African sample into the multigroup model. In this instance, two models are estimated: one incorporating strong and one incorporating weak factorial invariance. Note that the model utilizing weak factorial invariance is identical to the one with configural invariance due to the assumption of essentially τ-equivalent measures. The model incorporating all four countries and strong factorial invariance resulted in adequate model fit ($\chi^2 = 440.492$, $df = 135$, $p < .001$, RMSEA = 0.052, SRMR = 0.04, CFI = 0.904). As was the case in the initial sample of three countries, the CFI is rather low in this instance, but RMSEA and SRMR are well within the range of acceptable model fit (Hu & Bentler, 1999; Brown, 2015). As was indicated by Table 4.16, however, constraining the intercepts to be equal across countries does prove a bit too restrictive, with the LRT rejecting this assumption ($\chi^2 = 51.646$, $df = 5$, $p < .001$) and the information criteria unanimously favoring the less restrictive model (AIC: 104681.136 vs.

104639.49, BIC: 105447.283 vs. 105436.283, aBIC: 105050.1 vs. 105023.212).

Finally, the item selection performed for the other three countries results in a CFA with acceptable fit when applied only to the South African sample (the fit criteria are given in Table 4.11). Constructing a solution with the `stuart` approach, under the constraints described previously, resulted in a short-form overlapping with the one presented above in only three items. Notably, both items selected for Openness were also selected within the South African sample, as was Item 190, an indicator of the Conscientousness dimension. This model showed very good fit ($\chi^2 = 56.357$, $df = 30$, $p = 0.002$, RMSEA = 0.031, SRMR = 0.025, CFI = 0.975) within the South African sample.

While both selections provide well fitting measurement models in the South African sample, there are drastic differences in the reliabilities of the two-item dimensions. Predictably, the Openness dimension displays the same reliability in both solutions (0.527), and both Conscientiousness (0.525 vs. 0.584) and Extraversion (0.628 vs. 0.521) result in similar reliailities. For Agreeableness, however, the solution generated from the three other countries resulted in a reliability of 0.337, while the country specific solution provided a much better reliability at 0.592. On the opposite, the global solution provided a selection of Neuroticism items with a reliability of 0.702, while the specific solution resulted in just 0.531.

4.2.4 Discussion

This section presented an application of the `stuart` approach to a setting where a scale is constructed in a cross-cultural setting. Additionally, a complex combination of heuristic information is included in the search for a 10-item short scale to assess the Big Five.

While the final scale results in good model fit across multiple countries and shows adequate reliabilities for subscales consisting of just two items, the latent correlation structure is not satisfactory, as it is not in line with those found for other scales. As pointed out, this may be due to a selection resulting in the assessment of very specific sub-dimensions of the intended global dimensions. To remedy this, it is possible to include the latent correlation structure in the objective function and assign quality based on the deviation of the latent correlation matrix from the obejective matrix. Schultze and Eid (2015) presented the possibility of this by utilizing

$$\sum_{m=1}^{M} 1 - \min[(cor(m,m')_{\text{Orig.}} - cor(m,m')_s)^2, 1]$$

as part of the objective function. Of course, this approach should only be utilized if there is substantial knowledge about the appropriate latent correlations.

Overall, the results of this section do not speak for the appropriateness of a 10-item short scale consisting of items selected from a long-form with a large amount of specific sub-dimensions. While there is a possibility this may be limited to the 300 items of the IPIP presented here, it seems more beneficial to generate less specific questions for a more global scale, instead of selecting specific questions from a long form. Especially when considering model stability (or lack thereof) over a large number of possible solutions, a scale with just two two items assessing each sub-facet did not generate promising results.

Therefore, a conclusion from this application should be that either (*a*) a longer scale is necessary to assess the Big Five in meaningful manner, or (*b*) more general questions should be used in short forms for the superficial assessment of the Big Five. Both approaches have been applied in numerous variants. A third possibility - selecting six items, one per sub-dimension, for each of the five dimensions - does not seem promising, given the low correlations between such items that already lead to estimation problems in this application.

Nevertheless, this section showed a possibility of simultaneous item selection in multiple groups, thereby circumventing the need of multiple applications of item selection strategies in multiple groups. Instead, this approach allows for the aimed selection of items with relatively stable properties across different cultural settings.

4.3 Item Selection in MTMM Studies: Application in the Assessment of Emotional Expressivity

This section presents the item-selection using the `stuart` approach in a situation with multiple sources of information. As indicated in Section 2.3.4, MTMM structures are incorporated into the item-selection process via a restricted CTC(M-1) approach. In this approach, one source of measurement is defined as the reference (denoted m_u^*), and all other methods are contrasted against this method. This contrast is achieved via simple latent regression, allowing for the calculation of the consistency between the reference and all other methods.

In this application items are selected for the assessment of emotional expressivity. Gross and John (1997, p. 435) define emotional expressivity in such a way that an "individual is emotionally expressive to the extent that he or she manifests emotional impulses behaviorally". As such, emotional expressivity can be viewed as the observable behavior of an individual experiencing emotion, meaning this construct is not only focused on the person experiencing the emotion, but also on the social surroundings experiencing the reaction. Therefore there are (at least) two perspectives that must be taken into account when assessing emotional expressivity: the perspective of the person expressing emotion and the perspective of the people observing and experiencing the expression. Because of various sources of bias manifested in self- and informant-ratings (Lucas & Baird, 2006; Neyer, 2006) the overlap between the two is not necessarily high when assessing behaviors, especially when it is assessed via global, retrospective questionnaires. Therefore, it can be a more promising approach to assess both viewpoints to account for the specificities of the perspectives. If the assessment of emotional expressivity is intended via multiple sources of information, this should also be a part of scale construction and, therefore, be included in the process of item selection. Humrichouse (2010) presents an overview over the three most influential classic scales for the assessment of emotional expressivity, all of them originally constructed as self-reports but also used in the assessment of peer-reports. More recently, Trierweiler, Eid, and Lischetzke (2002) as well as Humrichouse (2010) presented scales intended for the use with self- and peer-reports, discussing the advantages of using these two complementary perspectives.

Beyond the question of the source of information, scales assessing emotional expressivity differ in their assumptions regarding the underlying dimensionality. While earlier scales were constructed under the assumption of multiple facets centering around a positive and a negative valence facet - the underlying assumption being that the emotional expressivity differs conceptually, depending on the valence of the experienced emotion (Gross & John, 1997) - more recent developments indicate more fine-grained, emotion-specific approaches (Trierweiler et al., 2002). Humrichouse (2010) combines these approaches in a hierarchical factor structure, with emotion-specific factors being subsumed under valence-specific factors, with a single dimension underlying these two. This hierarchical approach is intended to allow for some specificity at each level while simultaneously enabling the analysis of general relationships between emotional expressivity and other psychological constructs.

In this application, the goal is to select items for a short scale aimed at assessing the valence specific level in this hierarchy of facets. These items are selected from a pool of items assessing emotion-specific expressivity. Specifically, the data stem from the original study by Trierweiler et al. (2002), who assessed 482 individuals on 28 items using a self-report and two peer-reports. In direct comparisons of three different dimensionality assumptions, Trierweiler et al. (2002) concluded that the emotion-specific approach best represents the 24-item selection they made from the original 28 items. However, assessment of emotion-specific expressivity using 24 items may be more fine-grained than necessary in some applications, especially when the questionnaire is included in a study not primarily focused on this construct. Therefore, eight items (four per valence) will be selected with the hopes of creating a short-scale capable of a coarse assessment of emotional expressivity from both, a self- and a peer-report. To ensure the quality of this short scale for both possible sources, the MTMM approach described in Section 2.3.4 is utilized. To allow for validation, an approach similar to that employed in the previous two sections is used. The data include one self- and two peer-ratings per participant. The self- and one peer-report are used in item selection and the second peer-report is used as a hold-out for a validation of the quality achieved in this item selection.

4.3.1 Problem Representation

In line with the problem representations discussed this far, the problem of item selection is represented by the triple $(\mathcal{S}, f, \Omega)$, where $\mathcal{S}$ is the set of possible solutions, f is the objective

Table 4.17: Item allocation of the original 28 items assessing emotional expressivity.

Item	Word	Emotion	Valence	Item	Word	Emotion	Valence
1	affection	love	+	15	nervousness		-
2	joy	joy	+	16	cheerfulness	joy	+
3	fear	fear	-	17	concern	fear	-
4	anger	anger	-	18	fury	anger	-
5	shame	shame	-	19	regret	shame	-
6	sadness	sadness	-	20	sorrow	sadness	-
7	love	love	+	21	caring	love	+
8	happiness	joy	+	22	contentment	joy	+
9	worry	fear	-	23	anxiety	fear	-
10	resentment	anger	-	24	rage	anger	-
11	guilt	shame	-	25	embarassment	shame	-
12	depression	sadness	-	26	unhappiness	sadness	-
13	disgust		-	27	pride		+
14	intimacy	love	+	28	tension		-

function, and Ω is the set of constraints. Due to the dependency of the first on the last, the restrictions are discussed first.

As shown in Section 1.3.2, the fundamental constraints in all applications of the `stuart` approach are:

(ω_1) the sum of weights does not exceed capacity - Equation (1.10),

(ω_2) items are selected specifically within their facets - Equation (1.12), and

(ω_3) items may be selected to only one facet - Equation (1.13).

Due to the fact that the original 28 items do not differ substantially in the amount of time required to answer them (all 28 are single emotion words), all of the weights are set to $w_{im} = 1$. Additionally, as pointed out above, the goal is to construe a short scale with four items per valence-specific facet - thus, $a_m = 4$ for all facets. Note that this differs from the imbalance in the original scale used by Trierweiler et al. (2002), where 16 of the 24 items were allocated to the negative valence dimension. Of the 28 item pool used here, 19 are allocated to the negative dimension.

The constraints ω_2 and ω_3 are in line with substantive theory, whereby each of the 28 emotions has a clear positive or negative valence. For those unfamiliar with human emotions, Table 4.17 provides an overview of the items and their dimensional allocation. Items missing the emotional dimension were assessed in the study, but not used in the final analysis by Trierweiler et al. (2002).

Beyond these basic constraints, configural invariance is assumed across self- and peer-reports, meaning that the same items are selected for both informants. This has the benefit of generating a single, comparable short-form, but does not include overly restrictive assumptions (such as equality of item discrimination across sources), which are unlikely to hold for most items in the item pool. For both facets the self-report is chosen as reference. Note that, when using only two measurement methods and internal criteria, this selection is irrelevant, because the model implied covariance-matrix is identical, irrespective of the choice regarding the reference (Geiser et al., 2008). Thus, consistency coefficients, reliabilities, and model fit will be identical for both choices in this application. This, however, need not be the case in other applications using a more complex measurement model (i.e. models with more methods or external criteria).

More formally, the assumption of configural invariance defines two $\mathcal{B}_u$, one for each valence, containing two facets each. In this application the specific item-pools are denoted $\mathcal{C}_1$ for self-reported, positive valence items, $\mathcal{C}_2$ for self-reported, negative valence items, $\mathcal{C}_3$ for peer-reported, positive valence items, and $\mathcal{C}_4$ for peer-reported, positive valence items. Thus, $\mathcal{B}_1 := (1, 3)$ and $\mathcal{B}_2 := (2, 4)$. Because there are no repeated measures, each $\mathcal{B}_u$ translates directly into a $\mathcal{Q}_h$, constituting the fourth constraint (ω_4) imposed on the item selection in this application.

With $I_m = 9$ for $m \in \{1, 3\}$ (i.e. the positive valence facets), $I_m = 19$ for $m \in \{1, 3\}$ (i.e. the negative valence facets), and $a_m = 4$ as imposed by constraint ω_1, the set of possible solutions $\mathcal{S}$ is defined. This set contains a total of 488376 possible solutions[4].

In contrast to the applications presented in Sections 4.1 and 4.2, no heuristics are provided in this application. All items are assumed to be equally suitable for a short-form and no specific combinations are disallowed. Thus, pheromones are localized to nodes, as presented in Section 2.2.1. This is in line with the way the item selection was performed for the Ryff-Scale in Chapter 3.

To account for the MTMM nature of the setting, the consistency coefficient *con* is included in the objective function f. As pointed out above, this coefficient is akin to R^2 of the regression predicting the peer-reports by the self-reports. Therefore, it indicates convergent validity between the two sources of information. Because the aim of the short-form is to provide a general view of emotional expressivity, one of the aims in this selection is to maximize consistency between the different perspectives. This is done by including

$$\Phi_{con}(s) = \frac{1}{1 + e^{-10(M_{rel_{hs}} - .5)}} \tag{4.8}$$

as a component of the pheromone function. Here, $M_{con_{hs}}$ is the mean of consistencies across the two partitions $\mathcal{Q}_h$ in solution s. The top-left panel of Figure 4.5 shows this component pheromone function. The midpoint of this function is set to .5, because the agreement between self- and peer-ratings has been shown to be somewhere around .25 when using scales not explicitly optimized for consistency (Humrichouse, 2010). Because this is explicitly integrated into the objective function of the item-selection process in this application, the aim is set to double this consistency.

In addition to the consistency coefficient, RMSEA and SRMR are, again, included to indicate

[4] If the assumption of configural invariance were lifted, $\mathcal{S}$ would contain 238511117376 possible solutions.

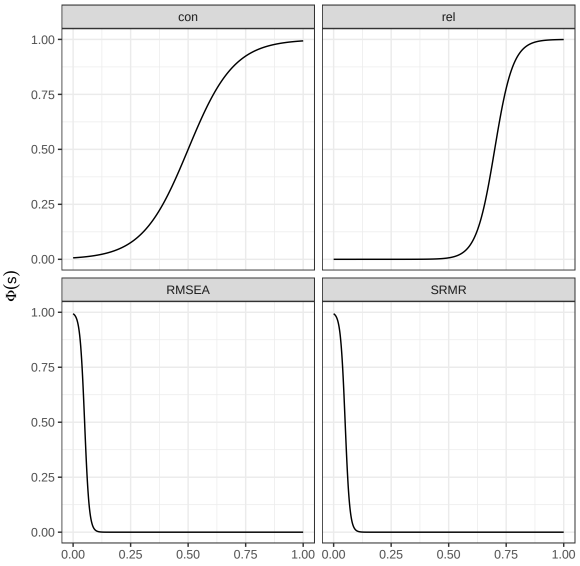


Figure 4.5: Component specific pheromone functions used in the application.

fit of the measurement model. The component specific pheromone functions are given by

$$\Phi_{\text{RMSEA}}(s) = 1 - \frac{1}{1 + e^{-100(\text{RMSEA}_s - .05)}} \tag{4.9}$$

$$\Phi_{\text{SRMR}}(s) = 1 - \frac{1}{1 + e^{-100(\text{SRMR}_s - .05)}} \tag{4.10}$$

and are also depicted in Figure 4.5.

The final component used to assess the quality of a solution is the reliability. As was done in the application in Section 4.2, the minimal reliability among the four facets is used to indicate reliability. This is done to ensure that the solution does not result in a short form which reliably assesses only three of the four facets but does not produce an acceptable measurement of the positive valence dimension in the self-report, for example. For the four-item, emotion-specific facets in the original 24 items, Trierweiler et al. (2002) reported reliabilities ranging from .63 to .88. Thus, .7 was chosen as the value of maximal discrimination in the reliability-based part of the fit function. This results in

$$\Phi_{rel}(s) = \frac{1}{1 + e^{-25(\min rel_{hs} - .7)}} \tag{4.11}$$

as the reliability component of the objective function. This function is also shown in Figure 4.5.

The components shown in Equations (4.8) to (4.11) are additively combined to form the overall pheromone function:

$$\Phi(s) = \Phi_{con}(s) + \Phi_{\mathrm{RMSEA}}(s) + \Phi_{\mathrm{SRMR}}(s) + \Phi_{rel}(s). \tag{4.12}$$

As was the case in previous applications, the objective function is defined as

$$f(s) = \begin{cases} \Phi(s), & \text{if } s \in \mathcal{S}^* \\ 0, & \text{else} \end{cases}, \tag{4.13}$$

with $\mathcal{S}^*$ being the set of viable solutions. All solutions resulting in non-positive-definite covariance matrices are assigned zero quality.

4.3.2 Item Selection

With the problem definition presented in the previous section, items were selected from the original pool using the parameter settings that were shown to most successful in Section 3.7.1. As noted above, no heuristics were used and pheromones were localized to nodes, in this application, making it very similar to the situation for which the performance of `stuart` was evaluated in Section 3.7.1. Thus, $\alpha = 1$, $\rho = .8$, $K = 16$, and the iteration-best deposit rule were used in this item selection. $K = 16$ was used instead of $K = 32$, because the problem is somewhat smaller than that evaluated in Chapter 3. In line with the recommendations made in Section 3.9, multiple instances of `stuart` (in this case five) were run to ensure the best possible selection. Appendix C.3 shows the R-code used to perform the `stuart` approach in this application.

In a simplified attempt to account for the non-normality of the variables in this application, the robust maximum-likelihood estimates based on the correction proposed by Satorra and Bentler (1994) and implemented in lavaan (Rosseel, 2012) were used. In this approach, parameter estimates are identical to those of a standard maximum-likelihood approach, but standard errors and the test-statistic are corrected, based on the mixture-distribution of the χ^2 test-statistic when data are non-normal. While standard errors are not of relevance in the construction of solutions or the determination of $f(s)$, the correction of the test-statistic also leads to a correction of the RMSEA (because this is based on the test-statistic). Estimates of the RMSEA tend to be biased in finite samples when data are non-normal, necessitating this correction when using the `stuart` approach with non-normal, continuous data. It should be noted, that alternatives to the approach by Satorra and Bentler (1994) have been proposed (e.g. Li & Bentler, 2006) and been shown to outperform the Satorra-Bentler approach (Brosseau-Liard, Savalei, & Li, 2012). However, as of writing, implementation of the correction based on the approach by Li and Bentler (2006) in lavaan was still work-in-progress.

The search required a total of 20704 CFAs to be run across all five replications (an average of 4140.8 per run). The best overall solution was found in two of the five instances. This solution selected the items “joy”, “happiness”, “cheerfulness”, and “contentment” to indicate the positive valence facet and “anger”, “resentment”, “fury”, and “rage” to indicate the negative facet, resulting in a $\Phi(s^{gb}) = 2.113$. This selection contains only items indicating the sub-facet “joy” from the original pool in place of the positive valence and only items indicating the “anger” sub-facet for the negative valence. This will be discussed in more detail in Section 4.3.4. The measurement model showed suboptimal fit in terms of RMSEA (0.062) and CFI (0.945), though both are close to the cut-off criteria stated by Hu and Bentler (1999). In terms of SRMR approximate model fit was adequate (0.04), though it should be noted that the SRMR is the only of these three fit criteria that is not corrected for non-normality, because it is not based on the test statistic. The model-test rejected the measurement model ($\chi^2 = 287.597$, $df = 100$, $p < .001$).

In terms of consistency between informants and reliability of the facets, the solution provides good results. For the positive facet, reliabilities were $rel_{\text{self}} = 0.827$ and $rel_{\text{peer}} = 0.719$ and for the negative facet they were $rel_{\text{self}} = 0.855$ and $rel_{\text{peer}} = 0.766$. Surprisingly, reliabilities were higher for self- than for peer-reports, indicating a more homogeneous assessment across the four respective items in the former. In contrast to the reliabilities, consistencies were higher for the positive than the negative facet (0.564 vs. 0.465), though both were adequate, achieving higher consistencies than those found for other scales (Humrichouse, 2010).

Due to the limited problem size (as pointed out in Section 4.3.1, $\mathcal{S}$ contains just 488376 possible solutions), it is feasible to determine the optimal solution under the constraints imposed in Ω via brute force. Investigating all possible solutions required approximately 27.211 hours utilizing R version 3.3.2 (R Core Team, 2016) and lavaan version 0.5-22 (Rosseel, 2012) on a machine with an Intel Core i7-5600U Quadcore CPU running Ubuntu 16.04. The optimal of these solutions is identical to the one recovered twice in five runs of the `stuart` approach. Of the 488376 solutions, 336129 (68.826%) were viable solutions, resulting in a pheromone different from 0. Note that, while the final solution did not result in satisfactory fit in terms of RMSEA and CFI, the minimal RMSEA in all possible solutions was 0.059 and the maximum CFI achieved was 0.947, indicating that satisfactory fit was not readily available for 8-item short scales constructed from the original pool in this sample.

Modification indices identified the residual correlation for the item “resentment” between the two informants as the most obvious detriment to model fit (MI = 72.361). Allowing for this residual correlation to indicate a specific effect of expressing “resentment” above and beyond the expression of the other selected negative valence emotions, results in a measurement model with acceptable model fit ($\chi^2 = 209.346$, $df = 99$, $p < .001$, RMSEA = 0.048, SRMR = 0.037, CFI = 0.968). Allowing for the uniqueness of this item had little effect on reliabilities and consistencies.

4.3.3 Selection Validation

To validate the item selection, the second peer-report is investigated closely in this section. In line with the procedure of the previous two applications, this validation will follow three main steps: (a) the method-specific models are investigated, (b) the third method is included in the overall model, and (c) the selection is compared to a selection performed only for the second peer rating. Due to the nature of the peer reports as being interchangeable sources of information in this data set (c.f. Eid et al., 2008), the second step also includes a series of tests for invariance among the peer-reports, to investigate whether the selection is suited for the use with multiple, interchangeable sources.

Figure 4.6 shows the path diagrams, with factor loadings and intercepts, of the three separate informant-specific models. The model for self-reports showed good overall fit ($\chi^2 = 28.909$, $df = 19$, $p = 0.067$, RMSEA = 0.033, SRMR = 0.024, CFI = 0.992), while those for the peer reports were both acceptable, albeit a bit worse. For the first peer-report, at 0.06, the RMSEA achieved the value defined as the cut-off by Hu and Bentler (1999). The values of the SRMR (0.024) and CFI (0.992) both indicated adequate model fit, though the χ^2-test rejected the model ($\chi^2 = 51.716$, $df = 19$, $p < .001$). For the second peer report, which was not included in the process of item selection, model fit was better than for the first ($\chi^2 = 44.856$, $df = 19$, $p = 0.001$, RMSEA = 0.053, SRMR = 0.032, CFI = 0.981). As indicated by the values shown in Figure 4.6, the structure of the measurement model was extremely similar between the two peer reports, while that for the self-reports differs a bit from the informant-reports. Self-reports show less latent variability and a somewhat different loading structure regarding the positive facet, though these differences appear rather small.

To allow for the test of equality in the measurement of the three informants, all three methods are included in one CFA, with the self-assessments being chosen as the reference method. The unconstrained model did not fit the data well ($\chi^2 = 689.538$, $df = 241$, $p < .001$, RMSEA = 0.062, SRMR = 0.044, CFI = 0.92) and modindices, again, indicated the necessity of allowing for specific effects concerning the item “resentment”. Allowing for the residual correlations of “resentment” across the three different sources dramatically improved model fit ($\Delta\chi^2_{SB} = 78.097$, $df = 3$, $p < .001$)[5].

Using the modified model as a baseline, Table 4.18 shows the fit and comparisons of the models incorporating the invariance assumptions described in Section 2.3.4. The sequential model comparisons using the Satorra-Bentler χ^2 (Satorra & Bentler, 2010) indicate that even the strict invariance assumption is adequate across the two peer-reports, showing support for the claim of interchangeability of these two informants. Imposing weak invariance between the two peer-reports and the self-report was rejected by the model comparison ($\Delta\chi^2_{SB} = 20.39$, $df = 6$, $p = 0.002$).

Performing the item-selection using only the data from the second peer-informants reveals a much different final selection of items.[6] In contrast to the selection made for the combination of

[5] Due to the correction of the test statistic, model comparisons are performed using the strictly positive Satorra-Bentler χ^2 (Satorra & Bentler, 2010)

[6] Performing the item-selection five times - as was done for the MTMM case - results in the exact same solution in all five cases.

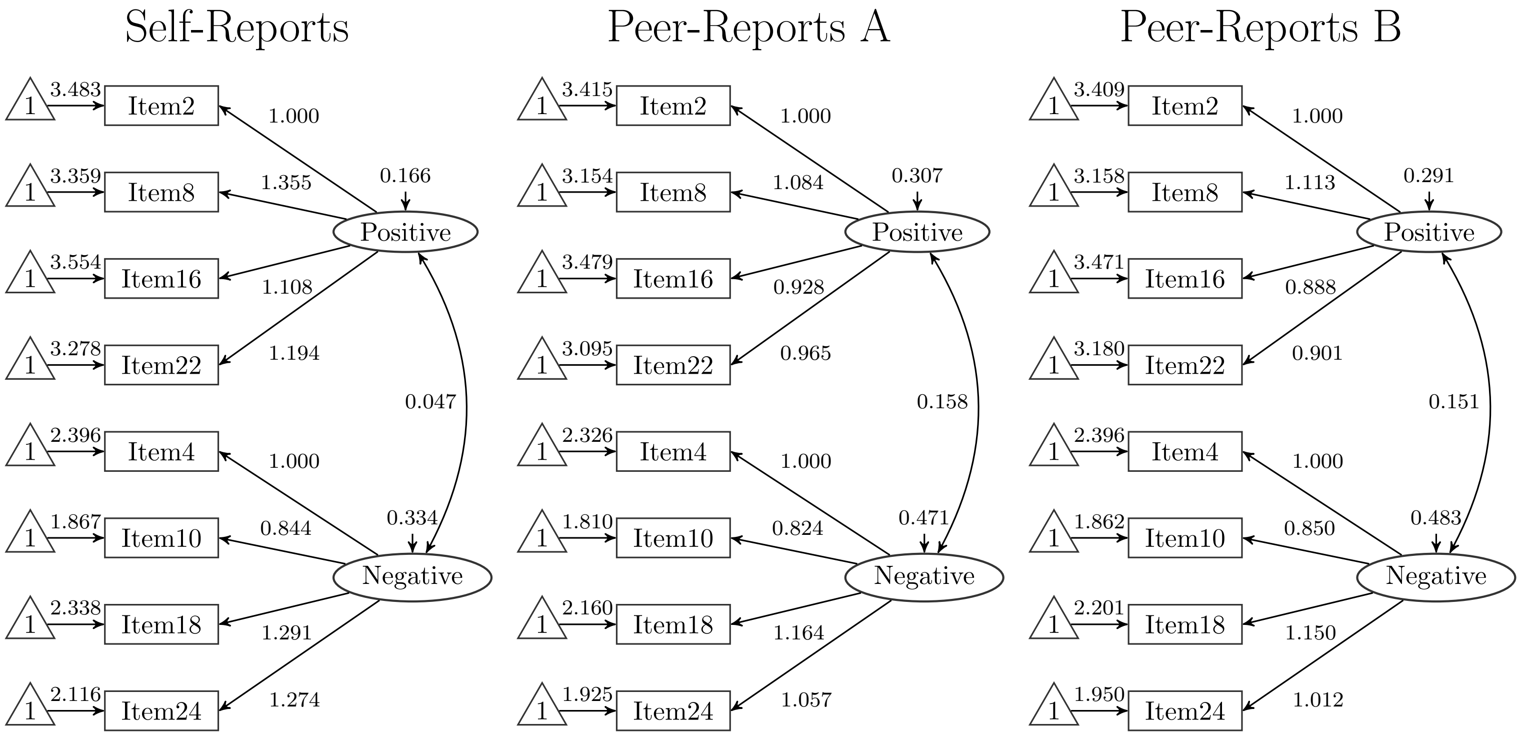


Figure 4.6: Path diagrams of the informant-specific models for the selected items.

Table 4.18: Model fit indices and comparisons of the models assuming different levels of invariance across the two peer reports.

	Configural	Weak	Strong	Strict
χ^2_{SB}	524.618	525.568	537.929	539.896
df	238	240	248	256
p	0.000	0.000	0.000	0.000
RMSEA	0.050	0.050	0.049	0.048
SRMR	0.040	0.038	0.039	0.039
CFI	0.949	0.949	0.948	0.949
AIC	20671.750	20666.060	20659.377	20648.267
BIC	21031.053	21017.008	20976.901	20932.368
aBIC	20758.096	20750.399	20735.684	20716.542
$\Delta\chi^2_{SB}$		0.400	10.927	3.918
Δdf		2	8	8
Δp		0.819	0.206	0.864

self-ratings and those provided by the first peer, the selection made for the second peer consists mainly of items indicating the facet “love” and only one indicating the facet “joy” - and thus only one item shared with the original selection. For the positive valence, the selection made, were items 1, 14, 21, and 22. For the negative valence, there was no overlap in selected items whatsoever, with items 3, 6, 9, and 20 being selected. Two of these indicate the emotion-facet “fear”, the other two the facet “sadness” in the original item pool. As pointed out above, the selection made originally, contained only anger items in the negative facet. This selection achieved almost perfect fit for the second peer-informants ($\chi^2 = 17.17$, $df = 19$, $p = 0.578$, RMSEA = 0, SRMR = 0.014, CFI = 1) in addition to the facets having high reliabilities (0.874 for positive and 0.86 for negative valence).

Beyond its qualities for the group of raters the selection was made in, this selection also shows good overall model fit when applied to the self-reports ($\chi^2 = 42.55$, $df = 19$, $p = 0.001$, RMSEA = 0.051, SRMR = 0.029, CFI = 0.974) and the peer-reports from the first group of peers ($\chi^2 = 35.522$, $df = 19$, $p = 0.012$, RMSEA = 0.042, SRMR = 0.021, CFI = 0.99) individually. This, however, underlines the importance of including multiple sources of information in the process of item selection, because the fit of the model when using this selection in the MTMM model was worse than that of the original selection ($\chi^2 = 771.307$, $df = 241$, $p = 0$, RMSEA = 0.068, SRMR = 0.046, CFI = 0.903) and consistencies were much worse (between .267 and .383).

4.3.4 Discussion

This section presented an application of the `stuart` approach to a situation with multiple sources of information. Using the parameter settings found to be most promising in Chapter 3, the optimal solution was found in two of five instances, requiring less than 5% of the runtime of the brute force approach (for all five instances combined).

While the solution fulfills the criteria set in the objective function (i.e. it provides a measurement model with acceptable fit, reliability, and consistency across self- and peer-ratings), the final selection does suffer from the typical reduction of construct content associated with an item selection based solely on statistical criteria. As can be expected, best reliabilities and consistencies were found for items indicating only two specific emotions contained in the original item pool. In this case, the short scale contains only words addressing the specific emotions joy and anger, resulting in a somewhat unsatisfactory collection for an assessment intended to globally assess positive and negative valence in emotional expressivity. This phenomenon can be avoided by implementing filter-type heuristics as shown in the previous two applications. However, this would result in a model with much lower values in terms of the objective function. As pointed out for the application to item selection from IPIP in Section 4.2, selecting items from a large pool of items designed to assess specific facets may not be a suitable approach to constructing short-scales intended for a more coarse assessment (perhaps at a higher level of facet abstraction). Instead, it may also be a better approach, in this case, to construct a pool of general items and select items for a short scale from this pool.

This application does underline the importance of using all sources of information when attempting to construct a scale intended for the use with different sources of information. Selecting items within the data specific to different perspectives results in solutions not suitable for the combined analysis of multiple perspectives. This shows one of the strengths of the `stuart` approach: selecting items in the MTMM is not any more difficult than item selection when using only one source of information. Using more classical approaches would require the integration of a number of different analytical procedures and their results. In this case, the results regarding the different informants would likely contradict each other and the selection of items may be immensely difficult. Using all information available in a single step of item selection is relatively simple within the `stuart` approach.

CHAPTER 5

Discussion

This thesis presented the `stuart` approach for item selection. Within this approach, the problem of selecting items is conceptualized more formally than in many previous strategies. Selecting items from a pool of items is understood as a combinatorial problem, specifically as an I-dimensional multiple knapsack problem with assignment restrictions (IMKAR), which can be solved by using adaptations of known algorithms. This problem conceptualization is formulated in such a way that it can flexibly be applied to any situation including one or more facets, groups, occasions, and sources of information. Thus, it constitutes a very inclusive framework for item selection that goes beyond previous approaches in its ability to be applied in such complex data situations without the need for the formulation of a new specific strategy for each single instance of item selection. By modeling the potential scales in confirmatory factor analysis (CFA), the `stuart` approach allows for a multitude of theoretically feasible measurement models and necessitates a clear formulation of the assumptions regarding the final scale's measurement structure. In contrast to many classical approaches for item selection, this allows for the selection of items which conform to a measurement structure which was defined a priori without requiring multiple phases of analyses. Additionally, the inclusion of complex study designs in a single step allows for the simultaneous optimization of the final scale regarding a plethora of different criteria - be they external, internal, or structural.

An ant colony optimization algorithm - an adaption of the $\mathcal{MAX} - \mathcal{MIN}$ Ant-System ($\mathcal{MM}$AS) algorithm presented by Stützle (1998) - is used to search for viable solutions within the space of combinations of items representing different possible scales. This algorithmic approach allows for the inclusion of pre-defined filtering (e.g. selecting anchor items and disallowing specific item combinations in the final scale). The performance of the search algorithm was investigated in an evaluation study, which culminated in a number of recommendations for applications, allowing for the use of the `stuart` approach without requiring any prior knowledge about the search algorithm itself. Beyond this, applications of the approach in a variety of different situations were presented in Chapter 4.

This chapter will focus on discussing the strengths and limitations of the `stuart` approach. First, attention will focus on the problem representation and the underlying conceptualization of item selection in this thesis. Section 5.2 discusses the chosen algorithm, before Section 5.3 deliberates the implications of the modeling approach. A section providing an overview of possible avenues for future extensions (Section 5.4) is followed by recommendations for applications (Section 5.5). This chapter will close with general conclusions about the `stuart` approach.

5.1 Conceptualization of Item Selection

As pointed out throughout Sections 1.3 and 2.1, the way item selection is conceptualized within the approach presented in this thesis is quite different from traditional approaches. While most traditional approaches are based on a rather informal problem definition and instead focus on deriving general recommendations based on prior experiences and the objectives (i.e. the quality of the final scale in terms of different aspects of validity, reliability, etc.), the `stuart` approach focuses on a more formal definition of the problem of item selection. This has the benefit of describing a minimal subset of concepts shared by all instances in which items are selected from a large pool to a form a scale. This problem definition is general enough to make a differentiation between item selection for short-forms and item selection for initial scale construction unnecessary. As pointed out in Chapter 1, once the item-pool is constructed, scale shortening and scale construction are so similar, that the same recommendations regarding the following steps apply to both of them (Smith et al., 2000). The same can be said about item selection for a questionnaire with a 5-point Likert-Scale and a questionnaire consisting of vignettes - the underlying problem can be conceived as the IMKAR described in Section 1.3, irrespective of the stimulus material. Such generality in the problem definition has the benefit of allowing for general solution approaches, which stay the same over a wide variety of instances, instead of requiring an instance specific approach for every single application.

The definition of the problem of item selection as an IMKAR has several specific advantages. First, it is a problem of known complexity. While this may seem like a minor victory, the implication is that any knowledge about problems of the same complexity also applies to the fundamental problem of item selection (as it is defined in this thesis). Second, it allows for the application - or at least adaptation - of several algorithmic strategies to the problem of item selection. Knapsack problems are wide-spread in many areas and therefore the fundamental approaches to solving them are established and evaluated. With a problem definition in place, approaches to solving KP can be adapted for the use in item selection, beyond the one approach presented in this thesis. Third, knowledge gained about item selection in the future can be transferred to other instances of item selection, if they are all conceptualized as IMKAR. Specifically, investigating the merits of other algorithmic approaches can have general implications for all specific applications of item selection. As long as the problem is defined as an IMKAR with its associated constraints, promising solution strategies can easily be applied to other instances and applications.

Conceptualizing item selection as an IMKAR is not without its drawbacks, however. As pointed out in Section 1.3, die IMKAR imposes three constraints:

$$a_m \geq \sum_{i=1}^{I} w_{im} x_{im} \tag{1.10, repeated}$$

$$\mathcal{C}_m^s \subseteq \mathcal{C}_m \tag{1.12, repeated}$$

$$\mathcal{C}_m^s \cap \mathcal{C}_{m'}^s \equiv \emptyset \qquad \forall m \neq m'. \tag{1.13, repeated}$$

The first, indicating that the sum of weights in a facet must not exceed the capacity of that facet, is perhaps the most problematic in the practice of item selection. Because a_m is part of the problem representation, it must be known a priori. While this is in line with the basic understanding of the `stuart` approach, that the substantive knowledge of a researcher should guide the construction of a scale, it may be too restrictive for some applications. It is possible to simply go through the automated process of item selection multiple times, for different values of a_m, but this would actually constitute a different problem, one containing multiple instances of the IMKAR. As such, solving this new, superordinate problem requires new strategies and approaches to find adequate solutions. This may sound overly complicated (why should one not simply search for an adequate solution with three, four, and five items per facet and pick the most promising one?), but it implies a much larger problem. Imagine a situation in which there are three facets and each $\mathcal{C}_m$ contains 20 items. Without imposing any restrictions, the meta-problem contains 20^3 possible item-number constellations. Each of these 8000 contains an IMKAR, resulting in a total 2.35×10^{17} possible solutions. Beyond the size of the problem, there is also the problem of interpretability. There are very few approaches to compare the psychometric quality of scales of different lengths in a meaningful fashion. Many CTT-based indicators of reliability are known to increase as the number of items increases (e.g. Raykov & Marcoulides, 2011), while model fit has been shown to decrease as the number of items increases (Moshagen, 2012).

The second and third constraints are closely related and have joint implications for applications. While the second constraint imposes that the facet-specific selection of items must come from a pre-defined set of items believed to be indicative of that facet, the second states that the final facet-specific selections must be disjoint sets. Strictly speaking, these two constraints do not necessarily require any prior knowledge of the measurement structure beyond the number of facets. It is possible to simply define all $\mathcal{C}_m \equiv \mathcal{C}$, bypassing any theory-guided allocation. At this point, the constraint imposed in Equation (1.13) will lead to a data-driven definition of facets. This may result in the same solutions being generated but not recognized as the same because the same sets of items are simply allotted to different facets - a phenomenon called label-switching in other analyses. However, the problem itself remains intact and even situations, in which nothing more than the number of theoretical facets is known, are included in the problem definition. The restriction of exclusivity, as imposed by Equation (1.13), prevents an item from being chosen twice, but the actual measurement structure must not be as rigid, because the inclusion of cross-loadings in a CFA-based approach is possible, or one could use other models classes such as EFA or exploratory SEM. Note, that within the `stuart` approach strict undimensionality is assumed, prohibiting the estimation of any cross-loadings, but this is not part of the problem definition itself.

To summarize: the problem definition presented in this thesis allows for a very inclusive representation, which encompasses all forms of item selection that are able to provide the number of facets being measured and the size of the final scale (e.g. number of items, time taken to complete the scale). Beyond this, the IMKAR imposes no limitations on the properties of any instance of item selection, making it extremely inclusive.

5.2 Algorithmic Approach

The second peculiarity of the approach to item selection presented in this thesis is the algorithmic approach to solving the IMKAR posited. The algorithm used for the `stuart` approach is an adaptation of the $\mathcal{MM}$AS algorithm proposed by Stützle (1998). As shown in Section 1.4, this is by no means the only ACO algorithm available, and the ACO metaheuristic is not the only class of algorithms suited to solve problems in the KP class.

Generally, a heuristic algorithm was chosen over a deterministic approach because of two main considerations: (a) the IMKAR representation of item selection can lead to excessively large problems and (b) when using CFA, a solution's quality can only be determined once an entire solution is constructed.

The first consideration precludes naïve deterministic approaches from being recommendable as a general solution strategy. In many cases, however, simple brute-force may be applicable to select items from the original pool. Because item selection needs to be performed only once - ideally - to construct a suitable candidate scale, it can be quite appropriate to run a brute-force search for several days. Because brute-force can easily be distributed to several cores, developments in computer hardware make these approaches more and more feasible in future applications. However, as described in Section 1.3, the complexity of the IMKAR is of $\mathcal{O}(n!)$ magnitude, meaning that the number of possible solutions increases rapidly with problem size. This means that larger item pools and more facets increase the overall runtime required for brute-force approaches immensely. Additionally, complex measurement models incorporating longitudinal data, multiple groups, and/or multiple sources of information increase the runtime of each single CFA. Beyond these specific cases discussed here, other factors, like missing data, can lead to increases in the time required for model estimation. In cases in which the estimation of a single model may require a minute, naïve deterministic algorithms quickly become unfeasible. For example, the item selection for the short-form of the Ryff-Scale required approximately 0.4 seconds per model during the evaluation shown in Chapter 3. Selecting three items per facet from the original 54-item scale gives 334569553920 possible solutions (see Section 3.3). Evaluating all possible solutions would therefore have required 4243.652 years of computational time (for reference, the pure computational time required for the evaluation presented in Chapter 3 was approximately 1581.999 days).

The second consideration precludes many of the remaining deterministic approaches to solving combinatorial problems. Most deterministic algorithms (for example the branch and bound strategies or tree-search algorithms) determine the quality of partial solutions to eliminate areas from the search space. However, if partial evaluation is not possible - as in the case of the IMKAR representation of item selection, see Equation (1.11) - these approaches cannot be applied without imposing additional assumptions about solution quality and solution construction.

Heuristic algorithms do not provide a guarantee of finding the best solution, however. This may be unsettling to researchers selecting items, but the basic idea of utilizing a heuristic algorithm is that some problems are too complex to solve with certainty, and thus they are solved in such a way that the solution fulfills the criteria set a priori. In this sense, the result of any single run of a `stuart`-based item selection does not necessarily provide the best solution, merely one deemed "good enough" in terms of the criteria. Additionally, different short forms can not only be found in different samples, but different runs of the approach on the same data may also lead to different results (as demonstrated by the evaluation presented in Chapter 3). Because the random components are not truly random, but performed by a computer - a deterministic machine by definition - providing random seeds makes selections reproducible within the same sample. Especially in light of the recent replication crisis in psychology and many other sciences (Open Science Collaboration, 2015), using such a heuristic approach may not seem too enticing at first glance. However, the central goal in an application of the `stuart` approach is the identification of a selection of items which fulfills the criteria imposed on the final scale by theory. Optimality of the solution, while nice, cannot be guaranteed when problems are too complex. Beyond this, final solutions are sample dependent to the degree in which the modeling approach and the criteria used in the objective function are sample dependent. This is the same in any other approach of item selection: a sample representative of the target population is required to ensure that the constructed scale is suited for use in this population.

Among the heuristic algorithmic approaches, the ACO metaheuristic was chosen because it has already shown promising results in the solving of several classes of KP (Alaya et al., 2004; Fidanova, 2003, 2007; Leguizamon & Michalewicz, 1999). In structural equation modeling, Marcoulides and Drezner (2003) proposed using ACO techniques for automated model specification searches and Leite et al. (2008) used ACO techniques for short-form construction. More recently, these approaches have been applied to different instances of item-selection and found to outperform classical and stepwise procedures (Janssen et al., 2015; Olaru et al., 2015; Schroeders et al., 2015). Applications of genetic algorithms (GA; Sahdra, Ciarrochi, Parker, & Scrucca, 2016) have also been shown to lead to promising results. While a comparison of the ACO and GA approaches favored GA (Schroeders et al., 2015), this is not indicative of the overall performance of ACO approaches, because of the ACO variant used.

Specifically, past approaches utilized a variant of the classical AS approach (Janssen et al., 2015; Leite et al., 2008; Olaru et al., 2015; Schroeders et al., 2015), which is known to have several shortcomings. As described in Section 1.4.2, the AS algorithm requires heuristic information to perform well (Dorigo & Stützle, 2004) and does not define convergence criteria, thereby running exactly as long as is imposed. None of the applications named utilized heuristic information and Schroeders et al. (2015) found the AS to perform as well as the GA variant used, albeit much slower, because it required more iterations. The $\mathcal{MM}$AS variant used in the `stuart` approach proposed in this thesis overcomes these shortcomings. The extensive evaluation presented in Chapter 3 shows that the `stuart` approach is able to retrieve optimal solutions, or construct solutions very close to optimal, in situations in which no heuristic information is provided. By imposing the convergence criteria presented in Equations (2.11) and (2.15) the runtime can be reduced and is made more independent of arbitrary settings for the number of maximum iterations (T). Additionally, Colorni et al. (1991) found AS to be unable to recover optimal solutions when the number of ants per colony (K) was chosen to be too small, irrespective of T. This is not the case with $\mathcal{MM}$AS and therefore also not for the `stuart` variant of it. As proven by Dorigo and Stützle (2004), $\mathcal{MM}$AS will find the optimal solution if given enough runtime, irrespective of the parameter settings (with the exception of requiring $\rho > 0$).

The `stuart` approach also goes beyond previous implementations in SEM automation by including multiple possibilities of pheromone localization. Even so, the presentation in this thesis is limited to the unidimensional and two-dimensional instances of pheromone deposit. The first is shown in Section 2.2.1 and is akin to localizing pheromones to the items themselves. In this way, items which were previously part of good solutions are more likely to be chosen into future solutions. The second is shown in Section 2.2.2 and localizes pheromones to the connections between two items. In this way, the pheromones affect the probability of choosing items dependent on the item that was previously chosen. It is possible to extend this localization to up to I dimensions. In this case, every complete solution would have an associated pheromone, making the process of constructing a solution simply a weighted random selection of s from $\mathcal{S}$, removing any component-based selection and thereby the very idea that leads to `stuart` being potentially quicker in finding good solutions than random search.

Table 5.1: Example of directed heuristics for the selection of items from specific sub-facets.

		One-on-One			Group			Public		
		c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
One-on-One	c_1	0	0	0	1	1	1	0	0	0
	c_2	0	0	0	1	1	1	0	0	0
	c_3	0	0	0	1	1	1	0	0	0
Group	c_4	0	0	0	0	0	0	1	1	1
	c_5	0	0	0	0	0	0	1	1	1
	c_6	0	0	0	0	0	0	1	1	1
Public	c_7	1	1	1	0	0	0	0	0	0
	c_8	1	1	1	0	0	0	0	0	0
	c_9	1	1	1	0	0	0	0	0	0

However, the limitation to two dimensions is rather strict, constituting a possible problem for a number of applications. Sections 4.1 and 4.2 show two applications which utilize the localization to arcs in order to impose logical filters in item selection. These filters represent substantive requirements, imposed upon the solutions by the researcher constructing the scale. In both cases these filters are also just two-dimensional (in the case of the MDBF, the number of positive and negative items was balanced in the final selection, and in the case of the IPIP, the two selected items were not allowed to stem from the same sub-dimension). Many applications may wish to impose logical gates that relate to more complex structures than the two-dimensional localization may permit at first glance.

Imagine a scale which assesses three specific dimensions of social anxiety. These three components are one-on-one situations, group activities, and public presentations. Each of these three specific components is assessed with three items, resulting in a scale with nine total items. The aim is to reduce this scale to three items, but it is necessary to ensure that each of the three items represents one of the three specific dimensions. This constraint can be imposed via the two-dimensional heuristics used in situations in which pheromones are localized to arcs. Because the heuristic matrices $\mathbf{H}_m$ do not need to be symmetric (in contrast to the pheromone matrices Φ_m, which are assumed to be symmetric, as pointed out in Section 2.2.2), a simple solution is to impose heuristics in a directed fashion to ensure that the items are always chosen in the order "one-on-one" → "group" → "public". Table 5.1 shows the implementation of these filter heuristics.

As stated in Section 2.2.2, the first item is chosen randomly. For any item, however, the $\mathbf{H}$ shown in Table 5.1, imposes that the next item that is chosen must be from the subsequent sub-facet. For example, selecting item 6 first (an item from the "group" sub-facet), results in

$$p(x_{(6,i')} = 1|t) = \frac{[\phi_{(6,i')}(t)]^{\alpha} 0^{\beta}}{\sum\limits_{i=1}^{9} [\phi_{(6,i)}(t)]^{\alpha} [\eta_{(6,i)}]^{\beta}} = 0 \qquad \forall i' \notin (7, 8, 9)$$

and

$$p(x_{(6,i')} = 1|t) = \frac{[\phi_{(6,i')}(t)]^{\alpha} 1^{\beta}}{\sum\limits_{i=1}^{9} [\phi_{(6,i)}(t)]^{\alpha} [\eta_{(6,i)}]^{\beta}} = \frac{[\phi_{(6,i')}(t)]^{\alpha}}{\sum\limits_{i'=7}^{9} [\phi_{(6,i')}(t)]^{\alpha}} \qquad \forall i' \in (7, 8, 9)$$

during the second selection, thereby guaranteeing that the second item is selected from the "public" sub-facet. Selecting any of these three items then guarantees that the final item must be in $(1, 2, 3)$ and therefore an item assessing the first sub-facet.

The approach used in this extremely simplified example can be used in every instance where the final solution must consist of equal numbers of items stemming from different sub-facets in the original item-pool. For example, the application shown in Section 4.2 can be extended to produce a short version of the IPIP-300 that consists of a total of 30 items, one selected from each sub-dimension within each of the five OCEAN dimensions. Note, however that this approach decreases problem size dramatically, possibly to the point where brute-force approaches are feasible. In the instance of the IPIP, this reduces the problem size from 3.15×10^{38} possible solutions, in the unrestricted variant, to 10^{30} when providing the filter heuristics. This is because the problem is actually reduced to a form where one out of ten items is selected for each of the 30 sub-facets, a rephrasing of the problem discussed in more detail in Section 5.3. While brute-force is not an option in this case, the feasibility of changing the algorithmic approach should be evaluated in every application imposing such wide-ranging restrictions via heuristics.

Beyond the possibility of using them as logical filters, heuristics should be understood as an initial aid to guide the search during early phases of the search process. This implies that heuristics must contain information which is beneficial to finding good solutions. While it is possible to intentionally steer the search in a direction which is favored based on theory, this should be done by including the substantive information in the objective function.

Beyond these specific points, utilizing an ACO-based algorithm to select items has a shortcoming embedded in its core approach to solution construction. While it was pointed out in Section 1.3 and discussed in Section 5.1 that the quality of components are not necessarily indicative of the overall quality of a solution consisting of these components, the use of pheromone deposits implies the utility of this assumption in the `stuart` approach. Pheromones are deposited to components stemming from good solutions, so that they may be selected with higher probability to other solutions. This only enhances the search speed (and solution quality under time constraints) if the assumption that components, which are part of good solutions are also more likely to be part of other good solutions, is true. Deterministic approaches were disregarded from consideration because of this very assumption, yet it is also imposed in ACO-based algorithms. One of the differences between deterministic and heuristic approaches is that the latter allow for this information to be wrong. An item may have been part of one good solution, but if it does not reappear in other good solutions its pheromone will evaporate and the fact that it once contributed to a good solution will soon be forgotten. Nevertheless, the basic assumption remains and, in item selection based on the results of CFA, it may be wrong. If the quality of solutions is not related to the quality of other solutions containing the same items, then the search space is completely unstructured and no search algorithm will consistently outperform random search. In such cases, the `stuart` approach may actually perform worse than random search, due to the pheromone deposits increasing the probability of reconstructing the same solution multiple times.

5.3 Modeling Approach

Section 2.3 shows the modeling approach utilized in this thesis for the evaluation of solutions created via the algorithm discussed in the previous section. As pointed out, a structural equation modeling and confirmatory factor analysis approach was chosen, allowing for the inclusion of complex data situations which may simultaneously include multiple groups, multiple measurement occasions, and multiple sources of information. The combination of these aspects for the process of item selection is shown in Section 2.3.5, where the facets are partitioned into the sets $\mathcal{Q}_h$, for which the same items must be selected. While this approach is more flexible in handling a multitude of data settings than other current approaches for item selection, it is not without its limitations. Perhaps the most apparent limitation lies in the handling of multiple groups and multiple occasions.

As pointed out in Section 2.3, cases in which items are selected for the use in multiple groups are handled via a multi-group CFA approach (e.g. Brown, 2015). Whether an application uses multiple groups or not, does not affect the item selection process itself, thereby imposing that $\mathcal{C}$, all $\mathcal{C}_m$, all $\mathcal{C}_m^s$, and the selection $\mathcal{C}_m \rightarrow \mathcal{C}_m^s$ are identical for all groups. This is a consequence of the multiple-group CFA being defined in the tradition of Jöreskog (1971), as the analysis of multiple covariance matrices of the same dimensions. While this may not be of substantial detriment to many instances of item selection, it precludes the `stuart` approach from being of use in situations with a stem of cross-culturally viable anchor items and additional culture-specific items. An extension to such cases is possible, albeit complicated, because it would most likely require the introduction of concepts from test-equating and linking into the `stuart` approach (see Dorans, Moses, & Eignor, 2010; W. Zhu, 1998, for an introduction to the matter). Further research is needed to extend the approach to such situations.

In cases with multiple occasions, the problem is similar, although easier to circumvent within the approach itself. As pointed out in Section 2.3, invariance assumptions are made on the basis of facets in the `stuart` approach. Thus, assuming configural (or any more restrictive) invariance over time, automatically assumes the equality of $\mathcal{C}_m$, $\mathcal{C}^s_m$, and the selection $\mathcal{C}_m \to \mathcal{C}^s_m$ for those m which are elements of the same $\mathcal{R}_v$. This equality is imposed in Equation (2.32) via the selection matrices $\boldsymbol{X}_m$. Full non-invariance can be imposed simply by not assigning the facets m and m' to the same $\mathcal{R}_v$. But this duality between full non-invariance and full configural invariance ignores all situations in which some items are used as anchors across occasions, while other items may be different across occasions. This may be necessary when scales are used to assess children over longer periods of time. As children age, the items used to assess them must also change, to ensure enough discrimination on the developing latent variable. As in the case of multiple groups, there is a vast collection of literature on test equating and linking in longitudinal settings (e.g. von Davier, 2011, for an overview). Extending the `stuart` approach to include these approaches presents an interesting avenue for future studies. Within the possibilities of the `stuart` approach in its current state, however, anchor items can be utilized by defining them as parts of repeated measures and their non-anchor counterparts as simply stemming from other pools of items.

Take a minimal example with three occasions and two sets of items (denoted $\mathcal{C}_1$ and $\mathcal{C}_2$) linking the first and the second occasion, as well as two sets of items linking the second and the third occasion (denoted $\mathcal{C}_3$ and $\mathcal{C}_4$). Additionally, there are occasion specific items ($\mathcal{C}_5$, $\mathcal{C}_6$, and $\mathcal{C}_7$). To ensure the same items are selected from the pool of items linking the first to the second occasion, $\mathcal{R}_1 = \{1, 2\}$ defines these two as repeated measures. $\mathcal{R}_2 = \{3, 4\}$ defines the two sets of items linking the second and third occasion as repeated measures. The partitions $\mathcal{Q}_h$ of the item pool are then simply $\{\mathcal{R}_1, \mathcal{R}_2, 5, 6, 7\}$. Attention must be paid to a_m to guarantee that the resulting scale has the same length at each occasion. To generate a scale with 20 items, for example, $a_1 = a_2 = 8$ ensures eight anchor items between the first and second occasion. This results in $a_5 = 12$, resulting in 20 items being selected to scale at the first occasion. For the links between the second and third occasion $a_3 = a_4 = 7$ may be chosen, resulting in $a_6 = 5$ and $a_7 = 13$. Thus, selecting items via `stuart` produces a solution with 20 items at each occasion and items linking the different occasions, while optimizing for the quality of their joint CFA model. Constraining the correlations between the latent variables representing facets assessed at the same occasion in the measurement model, leads to the assumption of unidimensionality for each of the occasion specific scales. In this way, the basic definitions of sets presented in Section 2.3.5 can be used to select items even in situations with more complex longitudinal designs.

This "trick" shows the flexibility of the set assignment strategy in the approach presented in this thesis. Few item-facet constellations, that are assumed a priori, cannot be expressed as a version of the three sets $\mathcal{C}_m$, $\mathcal{R}_v$, $\mathcal{B}_u$, and the resulting partitions $\mathcal{Q}_h$. The example used for the presentation of directed heuristics in Section 5.2 can also be rephrased to a problem with three facets, each with $a_m = 1$, instead of using directed heuristics. An application with single-indicator facets requires additional restrictions in the CFA, however, which is the reason the approach shown in Table 5.1 may be preferable.

Despite the flexibility in the formulation of facets and their relation to each other, the basic CFA models used to evaluate solutions are rather restrictive in their assumptions. As stated in Section 2.3.1, the implemented CFA models assume strict unidimensionality (i.e. cross-loadings of 0) and an absolute absence of residual correlations. The application presented in Section 4.3 for the item selection in a situation with multiple sources of information showed that this might be too strict in at least some applications. Especially in cases with repeated measures or a CTC(M-1) structure, the assumption of correlated residuals may often by suitable. The inclusion of these correlations is not possible for models generated by the `stuart` approach in an ad-hoc fashion, because any item may be part of any $\mathcal{C}_m$, leading to issues with model identification. Thus, it is necessary to state assumptions about violations of these assumptions prior to item selection.

The invariance assumptions imposed across groups, occasions, and possibly sources of information are also rather restrictive. While all invariance levels described by Meredith (1993) can be chosen at will, and invariance must not be assumed at all, contemporary approaches to partial and approximate measurement invariance are not included in this thesis. In contrast to the inclusion of general error-correlations or cross-loadings an extension to include these approaches is possible, however, because it does not systematically interfere with model identification. The inclusion of partial invariance is straightforward if invariance assumptions can be stated on an item-specific level a priori, and is, in fact, possible via rephrasing the facet structure in a similar fashion to that presented above for the case of scales with anchor items. Approximate measurement invariance can also be included rather simply on a theoretical level. However, including approximate measurement invariance currently presents a substantial practical problem. Because contemporary approaches to allow for slight inequalities of parameters in SEM and CFA are based on Bayesian estimation methods (e.g. B. O. Muthén & Asparouhov, 2013; van de Schoot et al., 2013), runtimes of an application of the `stuart` approach could be beyond acceptable.

Finally, the modeling approach implemented in this thesis is limited regarding the kinds of sources included in Section 2.3.4. Utilizing the CTC(M-1) approach implies that the sources of information are structurally different methods. These methods can be separated from interchangeable methods in the sense that they imply different random experiments (Eid et al., 2008; Koch, 2013; Nussbeck, Eid, Geiser, Courvoisier, & Lischetzke, 2009). Structurally different methods are generally assumed to differ in some substantial regard, thereby making them sources of somewhat different information regarding the subject. Interchangeable methods, on the other hand, are assumed to stem from a common population of methods, making them sources of the same information regarding the subject. Interchangeable methods can be modeled in multilevel SEM (Eid et al., 2008) or in a classic SEM (Nussbeck et al., 2009), but both approaches cannot to be recreated using the facet classification presented in Section 2.3. An extension to include these different types of methods is possible, but implies a substantial extension of the underlying system of facet sets in the `stuart` approach. Beyond this, the restricted CTC(M-1) approach (Geiser et al., 2008) implemented in the `stuart` approach is more prohibitive than strictly necessary when modeling structurally different methods. However, it allows relating facets to each other, instead of imposing item-level adaptations of the measurement models, making it more compatible with the flexibility of the facet definitions.

5.4 Extensions and Future Research

Some possible extensions were already touched upon in the previous sections, but this section is intended to provide an overview over the most promising avenues of extending the `stuart` approach beyond its current limitations, to further ensure its applicability as a method of item selection.

Perhaps the most global possible extension of the `stuart` approach is the inclusion of alternative heuristic algorithms. In the variant presented in this thesis, `stuart` is limited to the use of the adapted version of the $\mathcal{MM}$AS algorithm. As discussed above, Olaru et al. (2015) and Schroeders et al. (2015) found approaches based on genetic algorithms to be equally, if not more, viable for item selection. Integrating GA into the `stuart` approach would allow for a more direct comparison of the algorithms than these two studies showed, because both algorithms can be integrated into the same conceptualization of item selection.

Beyond this, Section 5.1 introduced the idea of extending the approach to include the length of the final scale as a variable in the optimization problem itself. In its current form, the approach presented in this thesis requires an a priori definition of a_m and all w_{im}, because they are constituents of the problem definition itself. They could also be reconceptualized as free parameters in a larger problem which must also be solved for scale length.

In terms of the measurement models included, there are several opportunities for future research to expand the approach. First, the models shown in Section 2.3 all pertain to continuous items. While many CFA and SEM applications to psychological scales use the assumption of continuous items as a simplification, with negligible bias regarding the estimation of latent constructs, scales using less than five possible response categories should be approached in a different fashion (Finney & DiStefano, 2013). Including models for ordinal and categorical items is an important step to increase the applicability of the `stuart` approach and towards an integration of IRT into the approach. Beyond this, an extension to include the concepts of scale linking and test equating, as eluded to in Section 5.3, also represents a possible extension of considerable merit.

Finally, the approach presented in this thesis requires a sample to respond to the entire item pool. While standard approaches to missing data are implemented in the SEM software used in the `stuart` implementation (see Enders, 2010, for an overview), a more thorough investigation of planned missingness in item selection via this approach is necessary. Especially in situations where extremely short scales are required (e.g. in experience sampling methods) assessing only few items in a realistic setting may be very beneficial for the external validity of the final scale.

5.5 Recommendations for Applications

Chapter 4 presented three specific applications and the accompanying difficulties and possible limitations. The most important thing to keep in mind is that the `stuart` approach is a method for item selection. As with all strategies to item selection, constructing a good scale is most likely when there is a sound theoretical basis for the construct, its assumed measurement structure, and the original pool of items. An item selection strategy, no matter how intricate, can only find the best possible combination of items in the original item pool. Thus, any application of the `stuart` approach must be preceded by substantive research to generate a pool of sensible items (either in form of a completely new pool of items or in the form of a validated long scale).

The item selection from the IPIP (see Section 4.2) and the selection for a scale to assess emotional expressivity (shown in Section 4.3) both highlight one of the potential pitfalls of scale shortening: the generality of the assessed facets is mirrored in the formulation of the items. Items constructed for the use in a long form that assesses specific sub-dimensions are likely to be phrased in such a way, that they mirror those specific areas of content. Unconstrained selection from a long form with the aim of constructing a short form, which assesses only the general, superordinate dimensions is likely to result in the selection of items pertaining to the same sub-dimension. Thus, instead of constructing a shorter scale, assessing constructs in more coarse fashion, this often leads to the reduction of the overall construct breadth. Generally, it may be more feasible to construct a new pool of less specific items and select a short form from these, than it is to construct a short form from a long version assessing more specific dimensions. If this is not possible or not deemed necessary, filter matrices should be used as heuristic information to ensure an adequate composition of the facets.

Perhaps the most important aspect of selecting items via the `stuart` approach is the definition of the objective function f. Because this is, intentionally, specific for each application only broad, general recommendations can be given. In most cases the objective function should contain components which assess the quality of a possible scale in different areas. It is important to keep in mind, that f determines what the scale is optimized for. If f contains only an estimate of reliability, then the search will very likely result in a solution with high reliability, but possibly abysmal model fit. In most cases, sensible components to include into the objective function can be divided into four categories.

The first category constitutes indicators of the quality of the assumed measurement model. Most of these belong to classical model fit criteria, such as the RMSEA, SRMR, CFI, etc. Optimizing for good model fit ensures that the measurement model, which is assumed to underly the final scale, is a suitable representation of the scale structure. Note that different model fit indexes evaluate model fit differently and are affected by different properties of each single application, thus necessitating the inclusion of multiple indexes in every application. However, model quality is not limited to fit, but can also encompass model adequacy. In all applications presented in this thesis, measurement models resulting in improper solutions were automatically rejected as unsuitable solutions. Thus, indicators in this category should be included in the objective function to ensure that the theoretical structure is actually present in the final scale selected.

The second category encompasses classical indicators of scale quality. Most notably, perhaps, an indicator of reliability may be suitable in many applications. Due to CFA being used to evaluate measurement models in the `stuart` approach, reliability measures based on these models are most appropriate (e.g. Yang & Green, 2010). Beyond simple indicators of reliability, i.e. estimating the reliability of each facet, indicators such as the reliability in the assessment of change can be utilized in some settings (Geiser, 2008) to specifically search for possible scales, which are well suited to assess change over time. Other classical indicators, such as the size of factor loadings or item difficulty can also be included in the objective function.

The third category includes indicators of structural validity beyond the measurement model. In cases with multiple sources of information the consistency coefficient represents such an indicator. Beyond this, the latent correlations can be compared to a specified value, thereby imposing assumptions about the relationship between the latent constructs in the process of item selection. Perhaps the situation in which this is of relevance most often, is when a short scale is constructed and the latent correlation matrix is optimized for its resemblance of the correlation matrix of the original version.

The fourth category consists of indicators of external validity. The `stuart` approach allows for the inclusion of external criteria in the CFA which is evaluated. Thus, correlations or regression weights linking the facets in the scale to external criteria can be used to maximize predictive validity, for example. On the other hand, correlations with other scales can be minimized to ensure discriminant validity of the scale for which the selection is being performed.

Irrespective of which components are included in the objective function, it is sensible to use a function with known limits. Most of the components listed above lie within pre-specified intervals, making this easier. Using an objective function for which the theoretical upper limit is known is beneficial, because it allows for an easier interpretation of solution quality and setting of the parameters discussed in Section 3.1. Beyond this, it is crucial to choose an objective function which is strictly positive. The quality of a (iteration-best or global-best) solution is deposited directly as pheromone in Φ, as shown in Equations (2.7) and (2.8). This pheromone is then used to determine the selection probability of an item via Equations (2.6) and (2.12). Using a quality function which may result in negative values, can lead to negative selection probabilities, resulting in failure of the application.

Beyond the quality function, the heuristics play an integral part in determining the performance of the `stuart` approach. Providing heuristics can have varying effects on runtime and solution quality and the in-depth discussion of the consequences of using heuristics beyond binary filters is given in Section 3.9. A general recommendation is to use heuristics as logical filters whenever necessary. In some cases, it may be more beneficial to rephrase the measurement model and include more specific facets, instead of defining global facets and disallowing combinations via heuristics, because these approaches allow for a more flexible handling of the violation of strict unidimensionality, which may result from such theory-driven constraints. Using more facets makes the measurement model more complex and flexible, while the imposition of heuristics regarding combinations of items makes the search problem itself more difficult. Beyond the logical filters, very large numbers can be used in $\mathbf{H}$ to guarantee selection of some items which are vitally important to the definition of the scale from a substantive perspective.

Recommendations regarding the specific settings of the algorithmic parameters α, β, ρ, K, the deposit-rule, as well as the possibility of using parameter schedules were derived from an extensive evaluation study and are given in Section 3.9.

5.6 Conclusion

The `stuart` approach presented in this thesis provides a flexible approach for item selection, either in the initial construction of scales or in the generation of short-forms of validated scales. The evaluation presented in Chapter 3 and the applications shown in Chapter 4 indicate that this approach is able to generate good solutions in a wide array of different situations. Utilizing a CFA approach to item selection allows for the a priori specification of the measurement model of the final scale. Due to the explicit inclusion of multiple groups, multiple measurement occasions, and multiple sources of information, the approach can be used to select items in complex study situations in a single step. The explicit requirement of a priori knowledge of the measurement model, the extent of the final scale, the required invariance restrictions, and an objective function allows basing each single application of the approach in the substantive theory surrounding the scale it is applied to. Beyond this substantive knowledge, however, an in-depth understanding of the underlying algorithm is not needed for the application of the `stuart` approach, because general recommendations regarding the algorithm's parameters are given in Section 3.9.

The approach is implemented in an R-Package, allowing for an application by anyone familiar with the basics of the statistical programming language R (R Core Team, 2016). This package is able to utilize both Mplus (L. K. Muthén & Muthén, 1998-2015) and lavaan (Rosseel, 2012) for the estimation of CFAs. Appendix C provides the R-Syntax required for the applications of the `stuart` approach presented in Chapter 4, for an easy introduction into the usage of the R-Package.

References

Abbott, R. A., Ploubidis, G. B., Huppert, F. A., Kuh, D., Wadsworth, M. E. J., & Croudace, T. J. (2006). Psychometric evaluation and predictive validity of Ryff's psychological well-being items in a UK birth cohort sample of women. *Health & Quality of Life Outcomes*, *4*, 76–16. doi: 10.1186/1477-7525-4-76

Afshar, M. (2009). Elitist-Mutated Ant System versus Max-Min Ant System: Application to pipe network optimization problems. *Scientia Iranica: Transaction on Civil Engineering*, *16*(4), 286-296. Retrieved from `http://en.journals.sid.ir/ViewPaper.aspx?ID=152533`

Ahmed, Z. H. (2013). A new reformulation and an exact algorithm for the quadratic assignment problem. *Indian Journal of Science and Technology*, *6*(4), 4368-4377. Retrieved from `http://www.indjst.org/index.php/indjst/article/view/31867`

Alaya, I., Solnon, C., & Ghédira, K. (2004). Ant algorithm for the multi-dimensional knapsack problem. In *International Conference on Bioinspired Optimization Methods and their Applications (BIOMA 2004)* (pp. 63–72). doi: 10.1.1.79.3439

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Belmont, CA: Wadsworth.

Anderson, C., & McShea, D. W. (2001). Individual versus social complexity, with particular reference to ant colonies. *Biological Reviews*, *76*(2), 211–237. doi: 10.1017/S1464793101005656

Baldacci, R., Hadjiconstantinou, E., & Mingozzi, A. (2003). An exact algorithm for the traveling salesman problem with deliveries and collections. *Networks*, *42*(1), 26–41. doi: 10.1002/net.10079

Beckers, R., Goss, S., Deneubourg, J. L., & Pasteels, J. (1989). Colony size, communication and ant foraging strategy. *Psyche*, *96*(3-4), 239–256. doi: 10.1155/1989/94279

Berger, S., & Freund, A. M. (2012). Fear of failure, disorganization, and subjective well-being in the context of preparing for an exam. *Swiss Journal of Psychology*, *71*(2), 83-91. doi: 10.1024/1421-0185/a000074

Birattari, M., & Dorigo, M. (2007). How to assess and report the performance of a stochastic algorithm on a benchmark problem: mean or best result on a number of runs? *Optimization Letters*, *1*, 309-311. doi: 10.1007/s11590-006-0011-8

Blum, C. (2005). Beam-ACO - hybridizing ant colony optimization with beam search: an application to open shop scheduling. *Computers & Operations Research*, *32*(6), 1565–1591. doi: 10.1016/j.cor.2003.11.018

Blum, C., & Roli, A. (2003). Metaheuristics in combinatorial optimization: Overview and conceptual comparison. *ACM Computing Surveys*, *35*(3), 268–308. doi: 10.1145/937503.937505

Blum, C., Roli, A., & Dorigo, M. (2001). HC-ACO: The hyper-cube framework for ant colony optimization. In *Proceedings of the MIC2001 – Metaheuristics International Conference, Vol. 2* (pp. 399–403). doi: 10.1.1.21.565

Bonabeau, E., Dorigo, M., & Theraulaz, G. (1999). *Swarm intelligence: From natural to artificial systems.* New York, NY, USA: Oxford University Press, Inc.

Boyle, G. J. (1991). Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? *Personality and Individual Differences*, *12*(3), 291–294. doi: 10.1016/0191-8869(91)90115-r

Brosseau-Liard, P. E., Savalei, V., & Li, L. (2012). An investigation of the sample performance of two nonnormality corrections for RMSEA. *Multivariate Behavioral Research*, *47*(6), 904-930. doi: 10.1080/00273171.2012.715252

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: Guilford Press.

Bullnheimer, B., Hartl, R. F., & Strauß, C. (1997). A new rank based version of the Ant System - a computational study. *Central European Journal for Operations Research and Economics*, *7*, 25–38. Retrieved from `http://epub.wu.ac.at/616/`

Burkard, R. E., Çela, E., Pardalos, P. M., & Pitsoulis, L. S. (1999). The quadratic assignment problem. In D.-Z. Du & P. M. Pardalos (Eds.), *Handbook of combinatorial optimization: Volume 1–3* (pp. 1713–1809). Boston, MA: Springer. doi: 10.1007/978-1-4613-0303-9_27

Burkard, R. E., & Offermann, J. (1977). Entwurf von Schreibmaschinentastaturen mittels quadratischer Zuordnungsprobleme [Design of typewriter keyboards via quadratic assignment problems]. *Zeitschrift für Operations Research*, *21*(4), B121–B132. doi: 10.1007/BF01918175

Burns, R. A., & Machin, M. A. (2009). Investigating the structural validity of Ryff's psychological well-being scales across two samples. *Social Indicators Research*, *93*(2), 359–375. doi: 10.1007/s11205-008-9329-1

Christofides, N., & Benavent, E. (1989). An exact algorithm for the quadratic assignment problem on a tree. *Operations Research*, *37*(5), 760–768. doi: 10.1287/opre.37.5.760

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309–319. doi: 10.1037//1040-3590.7.3.309

Clay Mathematics Institute. (2000). *Millenium problems.* `http://www.claymath.org/millennium-problems`. (Accessed: 2016-11-01)

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Colorni, A., Dorigo, M., & Maniezzo, V. (1991). Distributed optimization by ant colonies. In *Proceedings of the ECAL91 - European Conference on Artificial Life* (pp. 134–142). Paris, France: Elsevier.

Cordón, O., de Viana, I. F., & Herrera, F. (2002). Analysis of the Best-Worst Ant System and its variants on the QAP. In M. Dorigo, G. Di Caro, & M. Sampels (Eds.), *Ant Algorithms: Third International Workshop, ANTS 2002 Brussels, Belgium, September 12–14, 2002 Proceedings* (pp. 228–234). Berlin: Springer. doi: 10.1007/3-540-45724-0_20

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. doi: 10.1037/h0040957

Danner, D., Blasius, J., Breyer, B., Eifler, S., Menold, N., Paulhus, D. L., ... Ziegler, M. (2016). Current challenges, new developments, and future directions in scale construction. *European Journal of Psychological Assessment*, *32*(3), 175–180. doi: 10.1027/1015-5759/a000375

Dawande, M., Kalagnanam, J., Keskinocak, P., Ravi, R., & Salman, F. S. (2000). Approximation algorithms for the multiple knapsack problem with assignment restrictions. *Journal of Combinatorial Optimization*, *4*(2), 171–186. doi: 10.1023/A:1009894503716

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology*, *34*(4), 481–489. doi: 10.1037//0022-0167.34.4.481

Deneubourg, J., Aron, S., Goss, S., & Pasteels, J. (1990). The self-organizing exploratory pattern of the argentine ant. *Journal of Insect Behavior*, *3*(2), 159–168. doi: 10.1007/BF01417909

Deneubourg, J., Pasteels, J., & Verhaeghe, J. (1983). Probabilistic behavior in ants: A strategy of errors? *Journal of Theoretical Biology*, *105*, 259–271. doi: 10.1016/s0022-5193(83)80007-1

Di Caro, G., & Dorigo, M. (1998). AntNet: Distributed stigmergetic control for communications networks. *Journal Of Artificial Intelligence Research*, *9*(1), 317–365. doi: 10.1613/jair.530

Dinh, H. T., Mamun, A. A., & Dinh, H. T. (2005). Dynamically updating the exploiting parameter in improving performance of ant-based algorithms. In N. Megiddo, Y. Xu, & B. Zhu (Eds.), *Proceedings of Algorithmic Applications in Management: First International Conference, AAIM 2005, Xian, China, June 22-25, 2005.* (pp. 340–349). Berlin: Springer. doi: 10.1007/11496199_37

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The Mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment*, *18*(2), 192 - 203. doi: 10.1037/1040-3590.18.2.192

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). Principles and practices of test score equating. *ETS Research Report Series*, *2010*(2), 1–41. doi: 10.1002/j.2333-8504.2010.tb02236.x

Dorigo, M. (1992). *Optimization, learning and natural algorithms (in Italian)* (Unpublished doctoral dissertation). Dipartimento di Elettronica, Politecnico di Milano, Milan, Italy.

Dorigo, M., & Di Caro, G. (1999a). Ant colony optimization: a new meta-heuristic. In *Proceedings of the 1999 Congress on Evolutionary Computation* (Vol. 2, pp. 1470–1477). doi: 10.1109/CEC.1999.782657

Dorigo, M., & Di Caro, G. (1999b). New ideas in optimization. In D. Corne et al. (Eds.), (pp. 11–32). Maidenhead, UK: McGraw-Hill.

Dorigo, M., Di Caro, G., & Gambardella, L. M. (1999). Ant algorithms for discrete optimization. *Artificial Life*, *5*, 137–172. doi: 10.1162/106454699568728

Dorigo, M., & Gambardella, L. M. (1997). Ant Colony System: A cooperative learning approach to the traveling salesman problem. *IEEE Transactions on Evolutionary Computation*, *1*(1), 53–66. doi: 10.1109/4235.585892

Dorigo, M., Maniezzo, V., & Colorni, A. (1991). *The Ant System: An autocatalytic optimizing process* (Tech. Rep. No. 91-016). Politecnico di Milano.

Dorigo, M., Maniezzo, V., & Colorni, A. (1996). Ant system: optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *26*(1), 29–41. doi: 10.1109/3477.484436

Dorigo, M., & Stützle, T. (2004). *Ant colony optimization.* Cambridge, MA: The MIT Press.

Dorigo, M., & Stützle, T. (2010). Ant colony optimization: Overview and recent advances. In M. Gendreau & J.-Y. Potvin (Eds.), *Handbook of metaheuristics.* New York, NY: Springer. doi: 10.1007/978-1-4419-1665-5_1

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology*, *70*(4), 662 - 680. doi: 10.1037//0021-9010.70.4.662

Ducatelle, F. (2007). *Adaptive routing in ad hoc wireless multi-hop networks* (Unpublished doctoral dissertation). Universitá della Svizzera Italiana, Lugano.

Ehrhart, M. G., Ehrhart, K. H., Roesch, S. C., Chung-Herrera, B. G., Nadler, K., & Bradshaw, K. (2009). Testing the latent factor structure and construct validity of the Ten-Item Personality Inventory. *Personality and Individual Differences*, *47*(8), 900 - 905. doi: 10.1016/j.paid.2009.07.012

Eiben, A. E., Michalewicz, Z., Schoenauer, M., & Smith, J. E. (2007). Parameter control in evolutionary algorithms. In F. G. Lobo, C. F. Lima, & Z. Michalewicz (Eds.), *Parameter setting in evolutionary algorithms* (pp. 19–46). Berlin, Germany: Springer. doi: 10.1007/978-3-540-69432-8_2

Eiben, A. E., & Smith, J. E. (2015). *Introduction to evolutionary computing* (Second ed.). Berlin, Germany: Springer.

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika*, *65*(2), 241–261. doi: 10.1007/BF02294377

Eid, M., & Diener, E. (2004). Global judgments of subjective well-being: Situational variability and long-term stability. *Social Indicators Research*, *65*(3), 245–277. doi: 10.1023/B:SOCI

.0000003801.89195.bc

Eid, M., Geiser, C., Koch, T., & Heene, M. (2016). Anomalous results in G-factor models: Explanations and alternatives. *Psychological Methods*. doi: 10.1037/met0000083

Eid, M., Gollwitzer, M., & Schmitt, M. (2015). *Statistik und forschungsmethoden [Statistics and research methods]* (4th ed.). Basel, Switzerland: Beltz.

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CT-C(M-1) model. *Psychological Methods*, *8*(1), 38–60. doi: 10.1037/1082-989x.8.1.38

Eid, M., Nussbeck, F. W., Geiser, C., Cole, D. A., Gollwitzer, M., & Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: Different models for different types of methods. *Psychological Methods*, *13*(3), 230 - 253. doi: 10.1037/a0013219

Eid, M., & Schmidt, K. (2014). *Testtheorie und Testkonstruktion [Test theory and test construction]*. Göttingen, Germany: Hogrefe.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Escario, J. B., Jimenez, J. F., & Giron-Sierra, J. M. (2015). Ant colony extended: Experiments on the travelling salesman problem. *Expert Systems with Applications*, *42*(1), 390–410. doi: 10.1016/j.eswa.2014.07.054

Favaretto, D., Moretti, E., & Pellegrini, P. (2009). On the explorative behavior of Max–Min Ant System. In T. Stützle, M. Birattari, & H. H. Hoos (Eds.), *Proceedings of Engineering Stochastic Local Search Algorithms. Designing, Implementing and Analyzing Effective Heuristics: Second International Workshop, SLS 2009, Brussels, Belgium, September 3-4* (pp. 115–119). Berlin, Germany: Springer. doi: 10.1007/978-3-642-03751-1_10

Fernandes, H. M., Vasconcelos-Raposo, J., & Teixeira, C. M. (2010). Preliminary analysis of the psychometric properties of Ryff's scales of psychological well-being in Portuguese adolescents. *The Spanish Journal of Psychology*, *13*(2), 1032–1043. doi: 10.1017/s1138741600002675

Fidanova, S. (2003). ACO algorithm for MKP using various heuristic information. In I. Dimov, I. Lirkov, S. Margenov, & Z. Zlatev (Eds.), *Numerical Methods and Applications: 5th International Conference, NMA 2002 Borovets, Bulgaria, August 20–24, 2002 Revised Papers* (pp. 438–444). Berlin, Germany: Springer. doi: 10.1007/3-540-36487-0_49

Fidanova, S. (2007). Ant colony optimization and multiple knapsack problem. In J. P. Renard (Ed.), *Handbook of research on nature inspired computing for economics and management* (p. 498–509). Hershey, PA: Idea Group. doi: 10.4018/978-1-59140-984-7.ch033

Finch, J. (1987). The vignette technique in survey research. *Sociology*, *21*(1), 105–114. doi: 10.1177/0038038587021001008

Finney, S., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 439–492). Greenwich, CT: Information Age.

Fisher, R. A. (1924). On a distribution yielding the error functions of several well known statistics. *Proceedings of the International Congress of Mathematics, Toronto*(2), 805–813. Retrieved from `http://hdl.handle.net/2440/15183`

Fouskakis, D., & Draper, D. (2002). Stochastic optimization: a review. *International Statistical Review*, *70*(3), 315–349. doi: 10.1111/j.1751-5823.2002.tb00174.x

Fritz, T., Halfpaap, J., Grahl, S., Kirkland, A., & Villringer, A. (2013). Musical feedback during exercise machine workout enhances mood. *Frontiers in Psychology*, *4*, 921. doi: 10.3389/fpsyg.2013.00921

Fréville, A. (2004). The multidimensional 0–1 knapsack problem: An overview. *European Journal of Operational Research*, *155*(1), 1–21. doi: 10.1016/S0377-2217(03)00274-1

Furr, R. M. (2011). *Scale construction and psychometrics for social and personality psychology.* Los Angeles, CA: Sage Publications.

Gallo, G., Hammer, P. L., & Simeone, B. (1980). Quadratic knapsack problems. In M. W. Padberg (Ed.), *Combinatorial optimization* (pp. 132–149). Berlin: Springer. doi: 10.1007/BFb0120892

Gault, R. H. (1907). A history of the questionnaire method of research in psychology. *The Pedagogical Seminary*, *14*(3), 366–383. doi: 10.1080/08919402.1907.10532551

Geiser, C. (2008). *Structural equation modeling of multitrait-multimethod-multioccasion data* (Unpublished doctoral dissertation). Freie Universität Berlin, Berlin, Germany.

Geiser, C., Eid, M., & Nussbeck, F. W. (2008). On the meaning of the latent variables in the CT-C(M-1) model: a comment on Maydeu-Olivares and Coffman (2006). *Psychological methods*, *13*(1), 49–57. doi: 10.1037/1082-989X.13.1.49

Goldberg, L. R. (1972). Parameters of personality inventory construction and utilization: A comparison of prediction strartegies and tactics. *Multivariate Behavioral Research Monographs*, *72*(2), 5–59.

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological Assessment*, *4*(1), 26 - 42. doi: 10.1037//1040-3590.4.1.26

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*(1), 84–96. doi: 10.1016/j.jrp.2005.08.007

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528. doi: 10.1016/S0092-6566(03)00046-1

Goss, S., Aron, S., Deneubourg, J., & Pasteels, J. (1989). Self-organized shortcuts in the argentine ant. *Naturwissenschaften*, *76*(12), 579–581. doi: 10.1007/BF00462870

Gross, J. J., & John, O. P. (1997). Revealing feelings: Facets of emotional expressivity in self-reports, peer ratings, and behavior. *Journal of Personality and Social Psychology*, *72*(2), 435 - 448. doi: 10.1037//0022-3514.72.2.435

Guntsch, M., & Middendorf, M. (2002). A population based approach for ACO. In S. Cagnoni, J. Gottlieb, E. Hart, M. Middendorf, & G. R. Raidl (Eds.), *Proceedings of Applications of Evolutionary Computing: EvoWorkshops 2002: EvoCOP, EvoIASP, EvoSTIM/EvoPLAN Kinsale, Ireland, April 3–4* (pp. 72–81). Berlin, Germany: Springer. doi: 10.1007/3-540-46004-7_8

Hamann, M. (2015). *Essays on heuristic solution methods for combinatorial optimization problems* (Unpublished doctoral dissertation). Frankfurt School of Finance and Management, Mannheim.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston, MA: Kluwer-Nijhoff.

Hendriks, A., Hofstee, W. K., & Raad, B. D. (1999). The five-factor personality inventory (FFPI). *Personality and Individual Differences*, *27*(2), 307–325. doi: 10.1016/S0191-8869(98)00245-1

Hershberger, S. L. (2013). The problem of equivalent structural models. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 3–40). Greenwich, CT: Information Age.

Hölldobler, B., & Wilson, E. O. (1990). *The ants.* Cambride, MA: Belknap Press of Harvard University Press.

Hoos, H., & Stützle, T. (2004). *Stochastic local search: Foundations & applications.* San Francisco, CA: Morgan Kaufmann Publishers Inc.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. doi: 10.1080/10705519909540118

Humrichouse, J. J. (2010). *The hierarchical structure of emotional expressivity: scale development and nomological implications* (Doctoral dissertation, University of Iowa, Iowa City, IA). Retrieved from `http://ir.uiowa.edu/etd/519`

International Personality Item Pool: A scientific collaboratory for the development of advanced measures of personality traits and other individual differences. (2016). `http://ipip.ori.org/`. (Retrieved January 10, 2017)

Janssen, A. B., Schultze, M., & Grötsch, A. (2015). Following the ants: Development of short scales for proactive personality and supervisor support by ant colony optimization. *European Journal of Psychological Assessment*. doi: 10.1027/1015-5759/a000299

Ji, J., Song, X., Liu, C., & Zhang, X. (2013). Ant colony clustering with fitness perception and pheromone diffusion for community detection in complex networks. *Physica A: Statistical Mechanics and its Applications*, *392*(15), 3260–3272. doi: 10.1016/j.physa.2013.04.001

John, O., Donahue, E., & Kentle, R. (1991). *The Big Five Inventory-Versions 4a and 54.* Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.

Johnson, J. A. (2014). Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*, *51*, 78–89. doi: 10.1016/j.jrp.2014.05.003

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409–426. doi: 10.1007/bf02291366

Kafka, G. J., & Kozma, A. (2002). The construct validity of Ryff's scales of psychological well-being (SPWB) and their relationship to measures of subjective well-being. *Social Indicators Research*, *57*(2), 171–190. doi: 10.2307/27526988

Kállay, É., & Rus, C. (2014). Psychometric properties of the 44-item version of Ryff's psychological well-being scale. *European Journal of Psychological Assessment*, *30*(1), 15–21. doi: 10.1027/1015-5759/a000163

Ke, L., Feng, Z., Ren, Z., & Wei, X. (2010). An ant colony optimization approach for the multidimensional knapsack problem. *Journal of Heuristics*, *16*(1), 65–83. doi: 10.1007/s10732-008-9087-x

Kellerer, H., Pferschy, U., & Pisinger, D. (2004). *Knapsack problems.* Berlin, Germany: Springer.

Kennedy, J., & Eberhart, R. C. (2001). *Swarm intelligence.* San Francisco, CA: Morgan Kaufmann.

Kesebir, S. (2012). The superorganism account of human sociality: How and when human groups are like beehives. *Personality and Social Psychology Review*, *16*(3), 233–261. doi: 10.1177/1088868311430834

Kitamura, T., Kishida, Y., Gatayama, R., Matsuoka, T., Miura, S., & Yamabe, K. (2004). Ryff's psychological well-being inventory: Factorial structure and life history correlates among Japanese university students. *Psychological Reports*, *94*(1), 83–103. doi: 10.2466/pr0.94.1.83-103

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

Koch, T. (2013). *Multilevel structural equation modeling of multitrait-multimethod-multioccasion data* (Doctoral dissertation, Freie Universität Berlin, Berlin, Germany). Retrieved from `http://www.diss.fu-berlin.de/diss/receive/FUDISS_thesis_000000094645`

Kong, M., Tian, P., & Kao, Y. (2008). A new ant colony optimization algorithm for the multidimensional knapsack problem. *Computers & Operations Research*, *35*(8), 2672–2683. doi: 10.1016/j.cor.2006.12.029

Koopmans, T., & Beckmann, M. J. (1955). *Assignment problems and the location of economic activities* (Cowles Foundation Discussion Papers No. 4). Cowles Foundation for Research in Economics, Yale University. doi: 10.2307/1907742

Korn, C., Sharot, T., Walter, H., Heekeren, H., & Dolan, R. (2014). Depression is related to an absence of optimistically biased belief updating about future life events. *Psychological Medicine*, *44*(3), 579–592. doi: 10.1017/S0033291713001074

Kruyen, P. M. (2012). *Using short tests and questionnaires for making decisions about individuals: When is short too short?* Ridderker, The Netherlands: Ridderprint.

Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2012). Test length and decision quality in personnel selection: When is short too short? *International Journal of Testing*, *12*(4), 321–344. doi: 10.1080/15305058.2011.643517

Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2013). On the shortcomings of shortened tests: A literature review. *International Journal of Testing*, *13*(3), 223–248. doi: 10.1080/15305058.2012.703734

Laporte, G. (1992). The traveling salesman problem: An overview of exact and approximate algorithms. *European Journal of Operational Research*, *59*(2), 231–247. doi: 10.1016/0377-2217(92)90138-Y

Leguizamon, G., & Michalewicz, Z. (1999). A new version of ant system for subset problems. In *Proceedings of the 1999 Congress on Evolutionary Computation-CEC99* (Vol. 2, p. 1464). doi: 10.1109/cec.1999.782655

Leite, W. L., Huang, I.-C., & Marcoulides, G. A. (2008). Item selection for the development of short forms of scales using an ant colony optimization algorithm. *Multivariate Behavioral Research*, *43*(3), 411–431. doi: 10.1080/00273170802285743

Li, L., & Bentler, P. (2006). *Robust statistical tests for evaluating the hypothesis of close fit of misspecified mean and covariance structural models* (Tech. Rep.). Los Angeles, CA: Department of Statistics, UCLA. Retrieved from `http://escholarship.org/uc/item/4t29r830`

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*(140), 5–55.

Little, J. D., Murty, K. G., Sweeney, D. W., & Karel, C. (1963). An algorithm for the traveling salesman problem. *Operations Research*, *11*(6), 972–989. doi: 10.1287/opre.11.6.972

Little, T. D. (2013). *Longitudinal structural equation modeling.* New York, NY: Guilford Press.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*, 635–694. doi: 10.2466/PR0.3.7.635-694

López-Ibáñez, M., Stützle, T., & Dorigo, M. (2015). *Ant colony optimization: A component-wise overview* (Tech. Rep. No. TR/IRIDIA/2015-006). IRIDIA Université Libre de Bruxelles. Retrieved from `http://iridia.ulb.ac.be/IridiaTrSeries/`

Lord, F., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Welsley. doi: 10.2307/2283550

Lucas, R. E., & Baird, B. M. (2006). Global self-assessment. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 29–42). Washington, DC: American Psychological Association. doi: 10.1037/11383-003

Maniezzo, V. (1999). Exact and approximate nondeterministic tree-search procedures for the quadratic assignment problem. *INFORMS Journal on Computing*, *11*(4), 358–369. doi: 10.1287/ijoc.11.4.358

Maples, J. L., Guan, L., Carter, N. T., & Miller, J. D. (2014). A test of the international personality item pool representation of the revised NEO personality inventory and development of a 120-item IPIP-based measure of the five-factor model. *Psychological Assessment*, *26*(4), 1070–1084. doi: 10.1037/pas0000004

Marcoulides, G. A., & Drezner, Z. (2003). Model specification searches using ant colony optimization algorithms. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(1), 154–164. doi: 10.1207/S15328007SEM1001_8

Marsh, H. W., Ellis, L. A., Parada, R. H., Richards, G., & Heubeck, B. G. (2005). A short version of the Self Description Questionnaire II: Operationalizing criteria for short-form evaluation with new applications of confirmatory factor analyses. *Psychological Assessment*, *17*(1), 81–102. doi: 10.1037/1040-3590.17.1.81

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320–341. doi: 10.1207/s15328007sem1103_2

Mautor, T., & Roucairol, C. (1994). A new exact algorithm for the solution of quadratic assignment problems. *Discrete Applied Mathematics*, *55*(3), 281–293. doi: 10.1016/0166-218X(94)90014-0

McGuire, W. (1960). A syllogistic analysis of cognitive relations. In M. J. Rosenberh, V. I. Hovland, W. J. McGuire, R. P. Abelson, & J. W. Brehm (Eds.), *Attitude organization and change* (pp. 65–111). New Haven, CT: Yale University Press.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika*, *29*(2), 177–185. doi: 10.1007/BF02289699

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. doi: 10.1007/BF02294825

Millsap, R. E. (2011). *Statistical approaches to measurement invariance.* New York, NY: Routledge. doi: 10.4324/9780203821961

Moshagen, M. (2012). The model size effect in SEM: Inflated goodness-of-fit statistics are due to the size of the covariance matrix. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(1), 86–98. doi: 10.1080/10705511.2012.634724

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557–585. doi: 10.1007/BF02296397

Muthén, B. O., & Asparouhov, T. (2013). BSEM measurement invariance analysis. *Mplus Web Notes*, *17*. Retrieved from `https://www.statmodel.com/examples/webnotes/webnote17.pdf`

Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Neill, J. A., & Jackson, D. N. (1970). An evaluation of item selection strategies in personality scale construction. *Educational and Psychological Measurement*, *30*(3), 647–661. doi: 10.1177/001316447003000312

Neyer, F. J. (2006). Informant assessment. In M. Eid & E. Diener (Eds.), *Handbook of multi-method measurement in psychology* (p. 43-59). Washington, DC: American Psychological Association. doi: 10.1037/11383-004

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*(1), 1–18. doi: 10.1016/0022-2496(66)90002-2

Nussbeck, F. W., Eid, M., Geiser, C., Courvoisier, D. S., & Lischetzke, T. (2009). A CTC(M-1) model for different types of raters. *Methodology*, *5*(3), 88–98. doi: 10.1027/1614-2241.5.3.88

Olaru, G., Witthöft, M., & Wilhelm, O. (2015). Methods matter: Testing competing models for designing short-scale big-five assessments. *Journal of Research in Personality*, *59*, 56–68. doi: 10.1016/j.jrp.2015.09.001

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). doi: 10.1126/science.aac4716

Papadimitriou, C. H. (2003). Computational complexity. In A. Ralston, E. D. Reilly, & D. Hemmendinger (Eds.), *Encyclopedia of computer science* (4th ed., pp. 260–265). Chihcester, UK: John Wiley and Sons.

Pedemonte, M., Nesmachnow, S., & Cancela, H. (2011). A survey on parallel ant colony optimization. *Applied Soft Computing*, *11*(8), 5181–5197. doi: 10.1016/j.asoc.2011.05.042

Pellegrini, P., Favaretto, D., & Moretti, E. (2006). On MAX-MIN Ant System's parameters. In M. Dorigo, L. M. Gambardella, M. Birattari, A. Martinoli, R. Poli, & T. Stützle (Eds.), *Proceedings of ant colony optimization and swarm intelligence: 5th international workshop, ants 2006, brussels, belgium, september 4-7* (pp. 203–214). Berlin, Germany: Springer. doi: 10.1007/11839088_18

Pellegrini, P., Stützle, T., & Birattari, M. (2010). Off-line vs. on-line tuning: A study on MAX–MIN Ant System for the TSP. In M. Dorigo et al. (Eds.), *Proceedings of swarm intelligence: 7th international conference, ants 2010, brussels, belgium, september 8-10* (pp. 239–250). Berlin, Germany: Springer. doi: 10.1007/978-3-642-15461-4_21

Pellegrini, P., Stützle, T., & Birattari, M. (2012). A critical analysis of parameter adaptation in ant colony optimization. *Swarm Intelligence*, *6*(1), 23–48. doi: 10.1007/s11721-011-0061-0

Pisinger, D. (2007). The quadratic knapsack problem—a survey. *Discrete Applied Mathematics*, *155*(5), 623–648. doi: 10.1016/j.dam.2006.08.007

R Core Team. (2016). *R: A language and environment for statistical computing.* Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five inventory in english and german. *Journal of Research in Personality*, *41* (1), 203–212. doi: 10.1016/j.jrp.2006.02.001

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory.* New York, NY: Routledge.

Raykov, T., Marcoulides, G. A., & Li, C.-H. (2012). Measurement invariance for latent constructs in multiple populations: A critical view and refocus. *Educational and Psychological Measurement*, *72* (6), 954–974. doi: 10.1177/0013164412441607

Reingold, E. M., Nievergelt, J., & Deo, N. (1977). *Combinatorial algorthims: Theory and practice.* Englewood Cliffs, NJ: Prentice-Hall.

Risch, A. K., Strohmayer, C., & Stangier, U. (2005). *Psychologische Wohlbefindensskala - PWS.* (Unpublished Manuscript)

Romero, E., Villar, P., Gómez-Fraguela, J. A., & López-Romero, L. (2012). Measuring personality traits with ultra-short scales: A study of the Ten Item Personality Inventory (TIPI) in a Spanish sample. *Personality and Individual Differences*, *53* (3), 289–293. doi: 10.1016/j.paid.2012.03.035

Rosen, N. A., & Wyer, R. S. (1972). Some further evidence for the 'Socratic effect' using a subjective probability model of cognitive organization. *Journal of Personality and Social Psychology*, *24* (3), 420 - 424. doi: 10.1037/h0033665

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48* (2), 1–36. doi: 10.18637/jss.v048.i02

Ryff, C. D. (1989). Happiness is everything, or is it? Explorations on the meaning of psychological well being. *Journal of Personality and Social Psychology*, *57* (6), 1069–1081. doi: 10.1037//0022-3514.57.6.1069

Ryff, C. D., & Keyes, C. L. M. (1995). The structure of psychological well-being revisited. *Journal of Personality and Social Psychology*, *69* (4), 719–727. doi: 10.1037/0022-3514.69.4.719

Ryff, C. D., & Singer, B. H. (2006). Best news yet on the six-factor model of well-being. *Social Science Research*, *35* (4), 1103–1119. doi: 10.1016/j.ssresearch.2006.01.002

Sahdra, B. K., Ciarrochi, J., Parker, P., & Scrucca, L. (2016). Using genetic algorithms in a large nationally representative American sample to abbreviate the Multidimensional Experiential Avoidance Questionnaire. *Frontiers in Psychology*, *7*, 189. doi: 10.3389/fpsyg.2016.00189

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In *Latent variable analysis in developmental research* (p. 285-305). Thousand Oaks, CA: Sage Publications.

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference Chi-square test statistic. *Psychometrika*, *75* (2), 243–248. doi: 10.1007/s11336-009-9135-y

Schrijver, A. (2005). On the history of combinatorial optimization (till 1960). In G. N. K. Aardal & R. Weismantel (Eds.), *Discrete optimization* (Vol. 12, pp. 1–68). Paris, France: Elsevier.

doi: 10.1016/S0927-0507(05)12001-5

Schroeders, U., Wilhelm, O., & Olaru, G. (2015). Meta-heuristics in short scale construction: Ant colony optimization and genetic algorithm. *PLoS One*, *11*(11). doi: 10.1371/journal.pone.0167110

Schultze, M., & Eid, M. (2015). *Shortening Scales Based on Automated CFA Fitting.* (Meeting of the Working Group Structural Equation Modeling, Berlin (Germany), February 26-27)

Schumacker, R. E., & Lomax, R. G. (2016). *A beginner's guide to structural equation modeling* (4th ed.). New York, NY: Routledge, Taylor and Francis Group.

Simms, L. J. (2008). Classical and modern methods of psychological scale construction. *Social and Personality Psychology Compass*, *2*(1), 414–433. doi: 10.1111/j.1751-9004.2007.00044.x

Sirigatti, S., Penzo, I., Iani, L., Mazzeschi, A., Hatalskaja, H., Giannetti, E., & Stefanile, C. (2013). Measurement invariance of Ryff's psychological well-being scales across Italian and Belarusian students. *Social Indicators Research*, *113*(1), 67–80. doi: 10.1007/s11205-012-0082-0

Sirigatti, S., Stefanile, C., Giannetti, E., Iani, L., Penzo, I., & Mazzeschi, A. (2009). Assessment of factor structure of Ryff's psychological well-being scales in Italian adolescents. *Bollettino di Psicologia Applicata*, *259*(56), 30–50.

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment*, *12*(1), 102-111. doi: 10.1037//1040-3590.12.1.102

Springer, K. W., & Hauser, R. M. (2006). An assessment of the construct validity of Ryff's scales of psychological well-being: Method, mode, and measurement effects. *Social Science Research*, *35*(4), 1080–1102. doi: 10.1016/j.ssresearch.2005.07.004

Springer, K. W., Hauser, R. M., & Freese, J. (2006). Bad news indeed for Ryff's six-factor model of well-being. *Social Science Research*, *35*(4), 1120–1131. doi: 10.1016/j.ssresearch.2006.01.003

Stanton, J. M., Sinar, E. F., Balzer, W. K., & Smith, P. C. (2002). Issues and strategies for reducing the length of self-report scales. *Personell Psychology*, *55*(1), 167–194. doi: 10.1111/j.1744-6570.2002.tb00108.x

Steinmetz, H. (2013). Analyzing observed composite differences across groups. *Methodology*, *9*(1), 1–12. doi: 10.1027/1614-2241/a000049

Steyer, R. (1989). Models of classical psychometric test theory as stochastic measurement models: Representation, uniqueness, meaningfulness, identifiability, and testability. *Methodika*, *3*, 25–60.

Steyer, R., & Eid, M. (2001). *Messen und testen [Measuring and testing]* (2nd ed.). Heidelberg: Springer.

Steyer, R., Schwenkmezger, O., Eid, M., & Notz, P. (1991). *Befindlichkeitsmessung und Latent-State-Trait-Modelle [Mood measurement and latent-state-trait-models]* (No. DFG Projekt Ste 411/3-1). Retrieved from `http://www.metheval.uni-jena.de/materialien/ges7/`

`ZwischenberichtStufe1.php`

Steyer, R., Schwenkmezger, O., Notz, P., & Eid, M. (1997). *Der Mehrdimensional Befindlichkeitsfragebogen (MDBF) [The Multidimensional Mood State Questionnaire]*. Göttingen: Hogrefe.

Steyer, R., Schwenkmezger, O., Notz, P., & Eid, M. (2004). *Entwicklung des Mehrdimensionalen Befindlichkeitsfragebogens (MDBF). Primärdatensatz. [Development of the Multidimensional Mood State Questionnaire (MDBF). Primary data.].* Trier: Psychologisches Datenarchiv PsychData des Leibniz-Zentrums für Psychologische Information und Dokumentation (ZPID). doi: 10.5160/psychdata.srrf91en15

Stützle, T. (1998). *Local search algorithms for combinatorial problems: Analysis, improvements, and new applications* (Doctoral dissertation, Fachbereich Informatik, Universität Darmstadt, Darmstadt). Retrieved from `http://iridia.ulb.ac.be/~stuetzle/publications/Thesis.ThomasStuetzle.pdf`

Stützle, T., & Hoos, H. (2000). Max-Min Ant System. *Future Generation Computer Systems*, *16*(9), 889–914. doi: 10.1016/S0167-739X(00)00043-1

Stützle, T., López-Ibáñez, M., Pellegrini, P., Maur, M., Montes de Oca, M., Birattari, M., & Dorigo, M. (2010). *Paramter adaptation in ant colony optimization* (Tech. Rep. No. TR/IRIDIA/2010-002). IRIDIA Université Libre de Bruxelles. Retrieved from `http://iridia.ulb.ac.be/IridiaTrSeries/`

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408. doi: 10.1007/BF02294363

Tempel, K. M. E. (2016). *Deine Persönliche Glückswoche: Eine Untersuchung zur Effektivität von Interventionen auf Grundlage der positiven Psychologie zur Erhöhung des psychologischen Wohlbefindens* (Unpublished doctoral dissertation). Fachbereich Erziehungswissenschaft und Psychologie, Freie Universität Berlin, Berlin.

ten Holt, J. C., van Duijn, M. A. J., & Boomsma, A. (2010). Scale construction and evaluation in practice: A review of factor analysis versus item response theory applications. *Psychological Test and Assessment Modeling*, *52*(3), 272–297.

Torchiano, M. (2016). *effsize: Efficient effect size computation.* Retrieved from `https://CRAN.R-project.org/package=effsize` (R package version 0.6.2)

Trierweiler, L. I., Eid, M., & Lischetzke, T. (2002). The structure of emotional expressivity: Each emotion counts. *Journal of Personality and Social Psychology*, *82*(6), 1023–1040. doi: 10.1037//0022-3514.82.6.1023

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. doi: 10.1177/109442810031002

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & O., M. B. (2013).

Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology*, *4*(770), 1–15. doi: 10.3389/fpsyg.2013.00770

van de Schoot, R., Schmidt, P., & De Beuckelaer, A. (Eds.). (2015). *Measurement invariance.* Lausanne, Switzerland: Frontiers Media. doi: 10.3389/978-2-88919-650-0

van de Vijver, F., van Hermert, D., & Poortinga, Y. (Eds.). (2008). *Multilevel analysis of individuals and cultures.* London, UK: Psychology Press, Taylor & Francis Group.

van Dierendonck, D. (2004). The construct validity of Ryff's scales of psychological well-being and its extension with spiritual well-being. *Personality and Individual Differences*, *36*(3), 629–643. doi: 10.1016/s0191-8869(03)00122-3

van Dierendonck, D., Díaz, D., Rodríguez-Carvajal, R., Blanco, A., & Moreno-Jiménez, B. (2007). Ryff's six-factor model of psychological well-being, a Spanish exploration. *Social Indicators Research*, *87*(3), 473–479. doi: 10.1007/s11205-007-9174-7

Villar, F., Triadó, C., & Celdrán, M. (2010). Measuring well-being among Spanish older adults: Development of a simplified version of Ryff's scales of psychological well-being. *Psychological Reports*, *107*(1), 265–280. doi: 10.2466/02.07.08.10.21.PR0.107.4.265-280

von Baeyer, C. L., Chambers, C. T., & Eakins, D. M. (2011). Development of a 10-item short form of the Parents' Postoperative Pain Measure: The PPPM-SF. *The Journal of Pain*, *12*(3), 401–406. doi: 10.1016/j.jpain.2010.10.002

von Davier, A. A. (2011). *Statistical models for test equating, scaling, and linking.* New York, NY: Springer.

Wei, X. (2014). Parameters analysis for basic ant colony optimization algorithm in TSP. *International Journal of u- and e-Service, Science and Technology*, *7*(4), 159–170. doi: 10.14257/ijunnessst.2014.7.4.16

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). New York, NY: The Guilford Press.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research.* (pp. 281–324). Washington, DC: American Psychological Association. doi: 10.1037/10222-009

Wyer, R. S. (1974). Some implications of the "Socratic effect" for alternative models of cognitive consistency1. *Journal of Personality*, *42*(3), 399–419. doi: 10.1111/j.1467-6494.1974.tb00683.x

Yang, Y., & Green, S. B. (2010). A note on structural equation modeling estimates of reliability. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*(1), 66–81. doi: 10.1080/10705510903438963

Zhu, P., Zhao, M.-S., & He, T.-C. (2010). A novel ant colony optimization algorithm in application of pheromone diffusion. In K. Li, M. Fei, L. Jia, & G. W. Irwin (Eds.), *Proceedings of Life System Modeling and Intelligent Computing: International Conference on Life System Modeling and Simulation, LSMS 2010, and International Conference on Intelligent Computing for Sustainable Energy and Environment, ICSEE 2010, Wuxi, China, September 17-20* (pp. 1–8). Berlin, Germany: Springer. doi: 10.1007/978-3-642-15597-0_1

Zhu, W. (1998). Test equating: What, why, how? *Research Quarterly for Exercise & Sport*, *69*(1), 11–23. doi: 10.1080/02701367.1998.10607662

Ziegler, M., Kemper, C. J., & Kruyen, P. (2014). Short scales – five misunderstandings and ways to overcome them. *Journal of Individual Differences*, *35*(4), 185–189. doi: 10.1027/1614-0001/a000148

APPENDIX A

Abstracts

A.1 Abstract

Using questionnaires to assess constructs has a long standing tradition in psychological research. Several guidelines and best-practices for constructing questionnaires and scales have been proposed over the years. In most of these, it is recommended to generate more items than the final scale is supposed to include, test this item pool on a sample, and select those items that perform best for the (potentially) final scale. Recent developments have necessitated the use of much shorter scales, making the shortening of established scales a common setting in which items are selected from an original pool. Whether in scale shortening or in initial scale construction, the quality requirements for a valid and reliable scale are manifold and, not seldom, contradicting. Beyond this, modern psychological research is often based on complex study designs, making scales desirable, which are known to be adequate for longitudinal studies, multiple groups, multiple sources of information, or any combination thereof.

This thesis presents the `stuart` approach for item selection, which allows for the simultaneous consideration of a multitude of quality criteria in complex study settings. To this end, item selection is defined as an I-dimensional multiple knapsack problem with assignment restrictions (IMKAR) and an adaptation of the $\mathcal{MAX} - \mathcal{MIN}$ Ant-System ($\mathcal{MM}$AS) is presented as an algorithmic approach to find solutions for this problem. In this context, item selection is based on generating promising solutions for final scales, evaluating these solutions via confirmatory factor analysis (CFA), and using the results of these analyses to guide the search for better solutions. Within this approach, an ideal measurement model and its restrictions must be defined a priori and solutions are then generated to best accomplish this ideal. Utilizing the CFA approach allows for item selection based on measurement models including multiple facets, multiple occasions, multiple groups, and multiple sources of information simultaneously and optimizing the final solution for criteria of model fit under assumptions of measurement invariance, among others.

Because the aim of this thesis is to present an applicable, flexible approach for item selection, an extensive evaluation study was performed to investigate the performance of the chosen algorithmic approach and derive recommendations for applications. These recommendations were then transferred to three applications of item selection: (a) a longitudinal setting, incorporating measurement invariance over time as a crucial component in item selection for a mood scale, (b) a multiple-group setting, aimed at generating a cross-culturally comparable, ultra-short Big Five scale, (c) and a setting including self- and peer-reports in the step of item-selection, to generate a scale which can assess emotional expressivity via multiple sources of information.

Overall, the `stuart` approach proved flexible in the accommodation of a wide variety of study designs, allowing for complex, application-specific objective functions and measurement models. Additionally, the evaluation study allowed for the recommendation of parameter settings for the alogrithmic approach, which generated solutions very close to optimal.

A.2 Zusammenfassung

Das Verwenden von Fragebögen hat in der psychologischen Forschung eine lange Tradition. In diversen Richtlinien und Empfehlungen zur Fragebogenerstellung wird es empfohlen, mehr Items zu generieren als die finale Version des Fragebogens enthalten soll, diesen Item-Pool einer Stichprobe vorzulegen und dann die bestgeeigneten Items auszuwählen. Neuere Entwicklungen haben die Nutzung von viel kürzeren Skalen nötig werden lassen, sodass die Kürzung einer etablierten Skala ebenfalls zu einem gängigen Beispiel der Itemselektion geworden ist. Sowohl in Fällen der Skalenkürzung, als auch bei der Erstellung einer neuen Skala, sind die Qualitätsanforderungen an reliable und valide Skalen vielfältig und nicht selten widersprüchlich. Darüber hinaus ist moderne psychologische Forschung oft mit komplexen Studiendesigns verknüpft, wodurch Skalen vonnöten sind, welche für Längsschnittstudien, Multigruppenuntersuchungen, multi-methodale Studien oder eine Kombination aller drei geeignet sind.

In dieser Dissertation wird der `stuart` Ansatz vorgestellt, der die simultane Berücksichtigung diverser Qualitätskriterien in komplexen Studiendesigns bei der Itemselektion erlaubt. Dafür wird diese als I-dimensionales multiples Rucksackproblem mit Zuweisungsrestriktionen (IMKAR) definiert und eine Abwandlung des $\mathcal{MAX}-\mathcal{MIN}$ Ant-System ($\mathcal{MM}$AS) zu dessen Lösen präsentiert. In diesem Kontext werden Items dadurch selektiert, dass verschiedene, vielsprechende Lösungen generiert, via konfirmatorischer Faktorenanalyse (CFA) analysiert, und deren Ergebnisse für die Erstellung neuer Lösungen weiter verwendet werden. In diesem Ansatz wird ein idealisiertes Messmodell, mit all seinen Restriktionen, vorgegeben und Lösungen generiert, die dieses Ideal bestmöglich erfüllen sollen. Die CFA erlaubt es dabei, mehrere Facetten, Messzeitpunkte, Gruppen und Quellen von Information gleichzeitig in die Itemselektion einzuschließen und die Lösungen auf, beispielsweise, Modellpassungskriterien unter Invarianzannahmen zu optimieren.

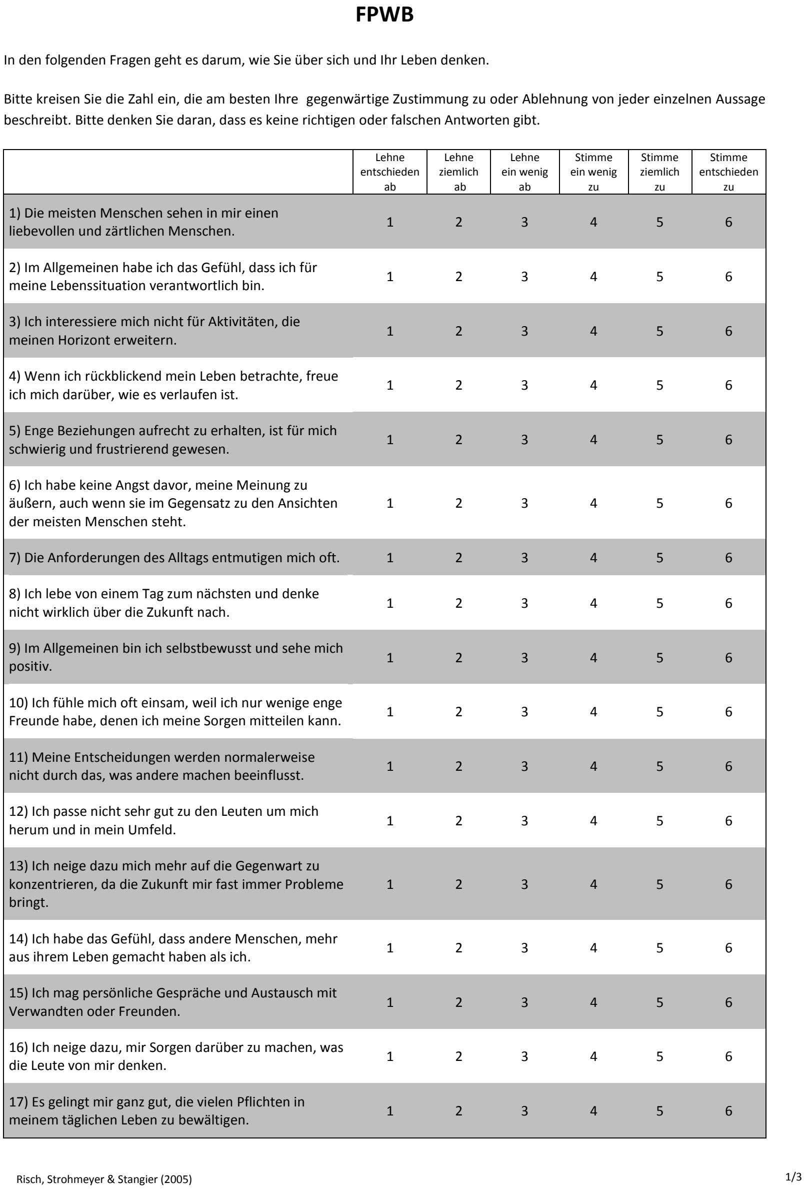
Da es das Ziel dieser Dissertation ist, einen anwendbaren und flexiblen Ansatz zur Itemselektion zu präsentieren, wurde eine extensive Evaluationsstudie durchgeführt, um das Verhalten des ausgewählten Algorithmus zu untersuchen und Empfehlungen für Anwendungen abzuleiten. Diese Empfehlungen wurden auf drei Anwendungen übertragen: (a) eine Längsschnittstudie, in der Messinvarianz eine wichtige Komponente in der Itemselektion für eine Wohlbefindensskala darstellt, (b) eine Multigruppenuntersuchung, in der eine Kurzskala für die interkulturell vergleichbare Erfassung der Big Five generiert werden soll und (c) eine Untersuchung, in der Selbst- und Fremdeinschätzungen in die Itemselektion einbezogen werden um eine Skala zur Erfassung von Emotionsausdruck zu erstellen.

Insgesamt erwies sich der `stuart` Ansatz als flexibel genug um die Itemselektion in einer Breite verschiedener Studiendesigns, unter Verwendung von anwendungsspezifischen Zielfunktionen und Messmodellen, zu ermöglichen. Zusätzlich konnten aus der Evaluationsstudie Parameterempfehlungen für den genutzten Algorithmus abgeleitet werden, welche Lösungen sehr nahe am Optimum generierten.

APPENDIX B

Questionnaires

B.1 Ryff-Scale



FPWB

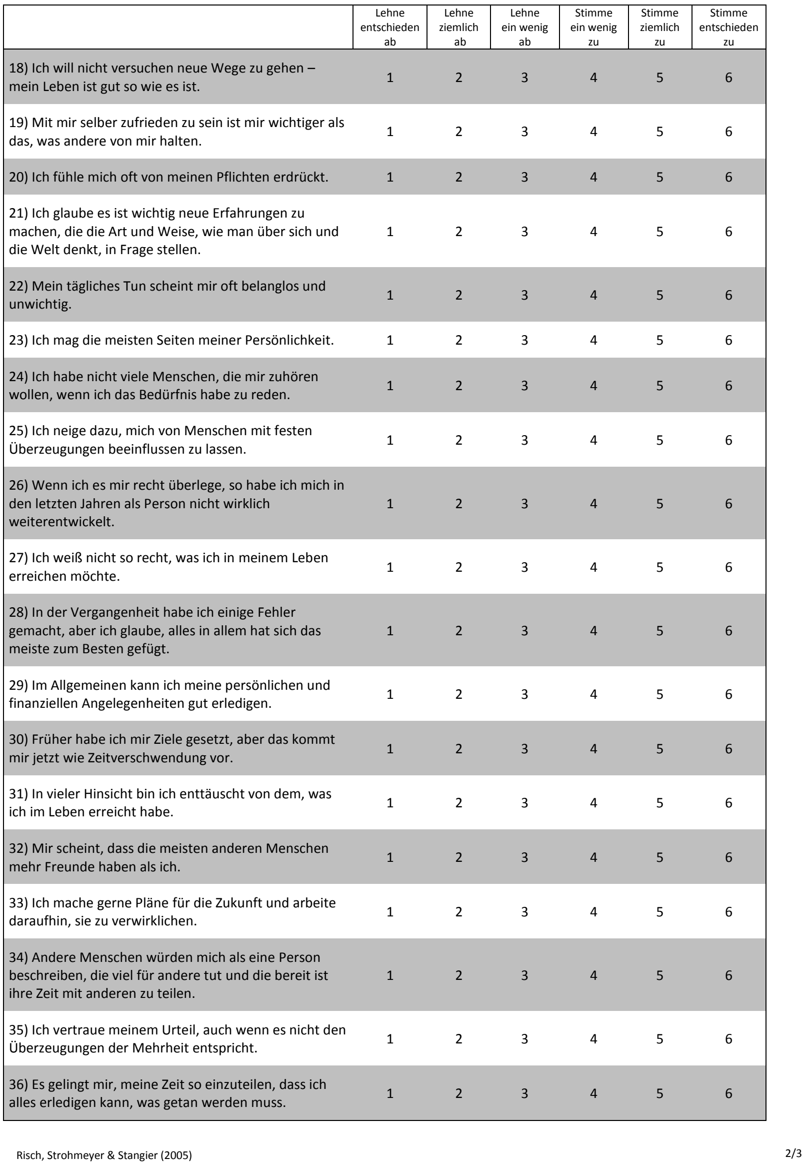
In den folgenden Fragen geht es darum, wie Sie über sich und Ihr Leben denken.

Bitte kreisen Sie die Zahl ein, die am besten Ihre gegenwärtige Zustimmung zu oder Ablehnung von jeder einzelnen Aussage beschreibt. Bitte denken Sie daran, dass es keine richtigen oder falschen Antworten gibt.

	Lehne entschieden ab	Lehne ziemlich ab	Lehne ein wenig ab	Stimme ein wenig zu	Stimme ziemlich zu	Stimme entschieden zu
1) Die meisten Menschen sehen in mir einen liebevollen und zärtlichen Menschen.	1	2	3	4	5	6
2) Im Allgemeinen habe ich das Gefühl, dass ich für meine Lebenssituation verantwortlich bin.	1	2	3	4	5	6
3) Ich interessiere mich nicht für Aktivitäten, die meinen Horizont erweitern.	1	2	3	4	5	6
4) Wenn ich rückblickend mein Leben betrachte, freue ich mich darüber, wie es verlaufen ist.	1	2	3	4	5	6
5) Enge Beziehungen aufrecht zu erhalten, ist für mich schwierig und frustrierend gewesen.	1	2	3	4	5	6
6) Ich habe keine Angst davor, meine Meinung zu äußern, auch wenn sie im Gegensatz zu den Ansichten der meisten Menschen steht.	1	2	3	4	5	6
7) Die Anforderungen des Alltags entmutigen mich oft.	1	2	3	4	5	6
8) Ich lebe von einem Tag zum nächsten und denke nicht wirklich über die Zukunft nach.	1	2	3	4	5	6
9) Im Allgemeinen bin ich selbstbewusst und sehe mich positiv.	1	2	3	4	5	6
10) Ich fühle mich oft einsam, weil ich nur wenige enge Freunde habe, denen ich meine Sorgen mitteilen kann.	1	2	3	4	5	6
11) Meine Entscheidungen werden normalerweise nicht durch das, was andere machen beeinflusst.	1	2	3	4	5	6
12) Ich passe nicht sehr gut zu den Leuten um mich herum und in mein Umfeld.	1	2	3	4	5	6
13) Ich neige dazu mich mehr auf die Gegenwart zu konzentrieren, da die Zukunft mir fast immer Probleme bringt.	1	2	3	4	5	6
14) Ich habe das Gefühl, dass andere Menschen, mehr aus ihrem Leben gemacht haben als ich.	1	2	3	4	5	6
15) Ich mag persönliche Gespräche und Austausch mit Verwandten oder Freunden.	1	2	3	4	5	6
16) Ich neige dazu, mir Sorgen darüber zu machen, was die Leute von mir denken.	1	2	3	4	5	6
17) Es gelingt mir ganz gut, die vielen Pflichten in meinem täglichen Leben zu bewältigen.	1	2	3	4	5	6

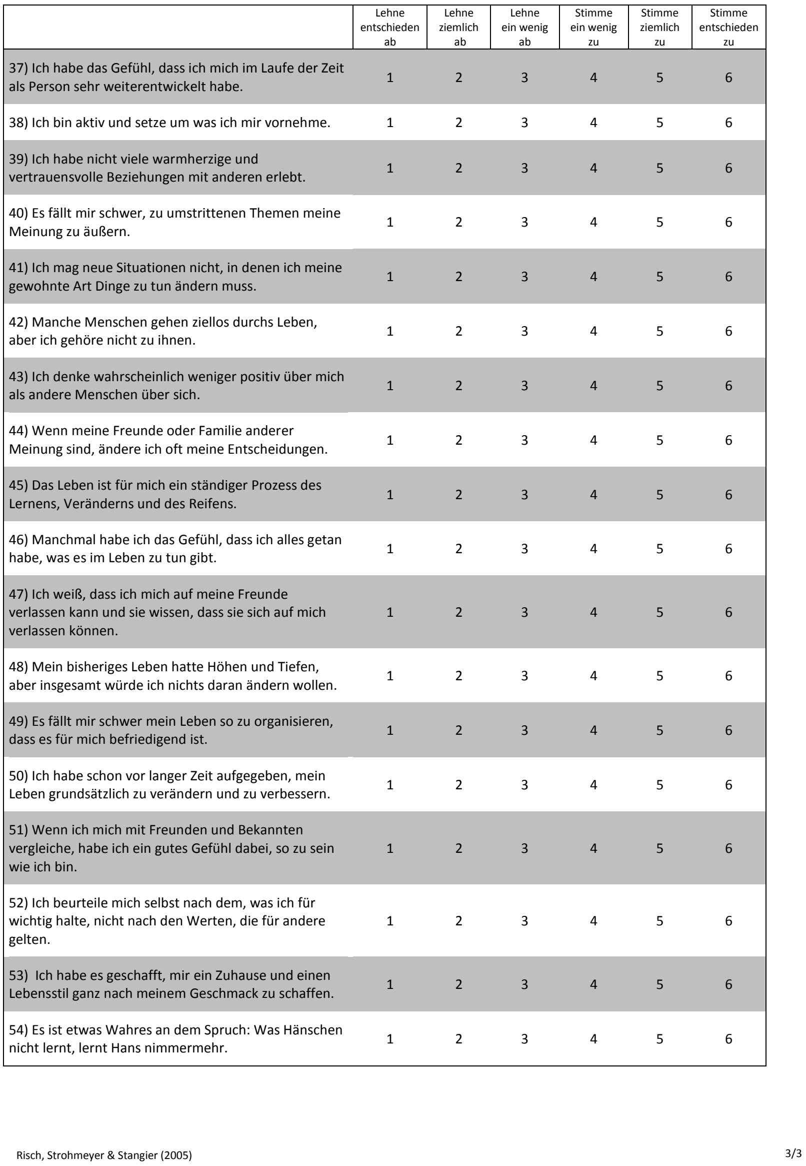
Risch, Strohmeyer & Stangier (2005)

1/3



	Lehne entschieden ab	Lehne ziemlich ab	Lehne ein wenig ab	Stimme ein wenig zu	Stimme ziemlich zu	Stimme entschieden zu
18) Ich will nicht versuchen neue Wege zu gehen – mein Leben ist gut so wie es ist.	1	2	3	4	5	6
19) Mit mir selber zufrieden zu sein ist mir wichtiger als das, was andere von mir halten.	1	2	3	4	5	6
20) Ich fühle mich oft von meinen Pflichten erdrückt.	1	2	3	4	5	6
21) Ich glaube es ist wichtig neue Erfahrungen zu machen, die die Art und Weise, wie man über sich und die Welt denkt, in Frage stellen.	1	2	3	4	5	6
22) Mein tägliches Tun scheint mir oft belanglos und unwichtig.	1	2	3	4	5	6
23) Ich mag die meisten Seiten meiner Persönlichkeit.	1	2	3	4	5	6
24) Ich habe nicht viele Menschen, die mir zuhören wollen, wenn ich das Bedürfnis habe zu reden.	1	2	3	4	5	6
25) Ich neige dazu, mich von Menschen mit festen Überzeugungen beeinflussen zu lassen.	1	2	3	4	5	6
26) Wenn ich es mir recht überlege, so habe ich mich in den letzten Jahren als Person nicht wirklich weiterentwickelt.	1	2	3	4	5	6
27) Ich weiß nicht so recht, was ich in meinem Leben erreichen möchte.	1	2	3	4	5	6
28) In der Vergangenheit habe ich einige Fehler gemacht, aber ich glaube, alles in allem hat sich das meiste zum Besten gefügt.	1	2	3	4	5	6
29) Im Allgemeinen kann ich meine persönlichen und finanziellen Angelegenheiten gut erledigen.	1	2	3	4	5	6
30) Früher habe ich mir Ziele gesetzt, aber das kommt mir jetzt wie Zeitverschwendung vor.	1	2	3	4	5	6
31) In vieler Hinsicht bin ich enttäuscht von dem, was ich im Leben erreicht habe.	1	2	3	4	5	6
32) Mir scheint, dass die meisten anderen Menschen mehr Freunde haben als ich.	1	2	3	4	5	6
33) Ich mache gerne Pläne für die Zukunft und arbeite daraufhin, sie zu verwirklichen.	1	2	3	4	5	6
34) Andere Menschen würden mich als eine Person beschreiben, die viel für andere tut und die bereit ist ihre Zeit mit anderen zu teilen.	1	2	3	4	5	6
35) Ich vertraue meinem Urteil, auch wenn es nicht den Überzeugungen der Mehrheit entspricht.	1	2	3	4	5	6
36) Es gelingt mir, meine Zeit so einzuteilen, dass ich alles erledigen kann, was getan werden muss.	1	2	3	4	5	6

Risch, Strohmeyer & Stangier (2005) 2/3

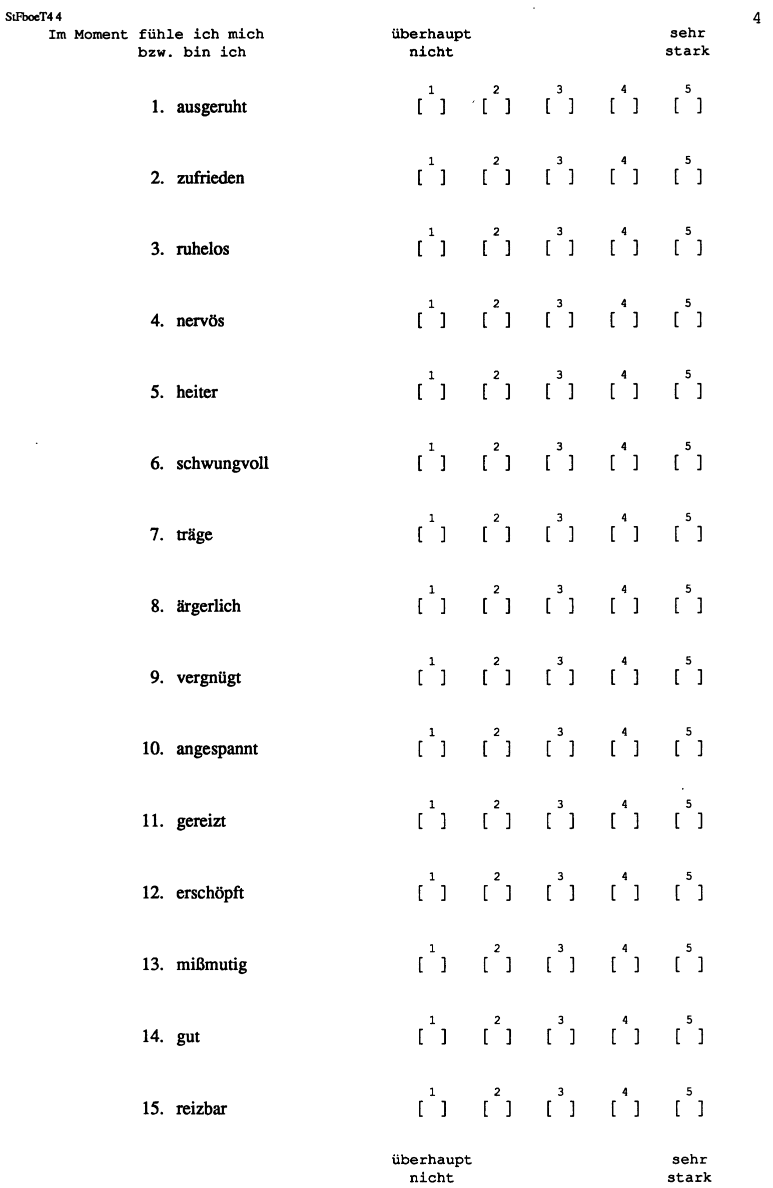


	Lehne entschieden ab	Lehne ziemlich ab	Lehne ein wenig ab	Stimme ein wenig zu	Stimme ziemlich zu	Stimme entschieden zu
37) Ich habe das Gefühl, dass ich mich im Laufe der Zeit als Person sehr weiterentwickelt habe.	1	2	3	4	5	6
38) Ich bin aktiv und setze um was ich mir vornehme.	1	2	3	4	5	6
39) Ich habe nicht viele warmherzige und vertrauensvolle Beziehungen mit anderen erlebt.	1	2	3	4	5	6
40) Es fällt mir schwer, zu umstrittenen Themen meine Meinung zu äußern.	1	2	3	4	5	6
41) Ich mag neue Situationen nicht, in denen ich meine gewohnte Art Dinge zu tun ändern muss.	1	2	3	4	5	6
42) Manche Menschen gehen ziellos durchs Leben, aber ich gehöre nicht zu ihnen.	1	2	3	4	5	6
43) Ich denke wahrscheinlich weniger positiv über mich als andere Menschen über sich.	1	2	3	4	5	6
44) Wenn meine Freunde oder Familie anderer Meinung sind, ändere ich oft meine Entscheidungen.	1	2	3	4	5	6
45) Das Leben ist für mich ein ständiger Prozess des Lernens, Veränderns und des Reifens.	1	2	3	4	5	6
46) Manchmal habe ich das Gefühl, dass ich alles getan habe, was es im Leben zu tun gibt.	1	2	3	4	5	6
47) Ich weiß, dass ich mich auf meine Freunde verlassen kann und sie wissen, dass sie sich auf mich verlassen können.	1	2	3	4	5	6
48) Mein bisheriges Leben hatte Höhen und Tiefen, aber insgesamt würde ich nichts daran ändern wollen.	1	2	3	4	5	6
49) Es fällt mir schwer mein Leben so zu organisieren, dass es für mich befriedigend ist.	1	2	3	4	5	6
50) Ich habe schon vor langer Zeit aufgegeben, mein Leben grundsätzlich zu verändern und zu verbessern.	1	2	3	4	5	6
51) Wenn ich mich mit Freunden und Bekannten vergleiche, habe ich ein gutes Gefühl dabei, so zu sein wie ich bin.	1	2	3	4	5	6
52) Ich beurteile mich selbst nach dem, was ich für wichtig halte, nicht nach den Werten, die für andere gelten.	1	2	3	4	5	6
53) Ich habe es geschafft, mir ein Zuhause und einen Lebensstil ganz nach meinem Geschmack zu schaffen.	1	2	3	4	5	6
54) Es ist etwas Wahres an dem Spruch: Was Hänschen nicht lernt, lernt Hans nimmermehr.	1	2	3	4	5	6

Risch, Strohmeyer & Stangier (2005)

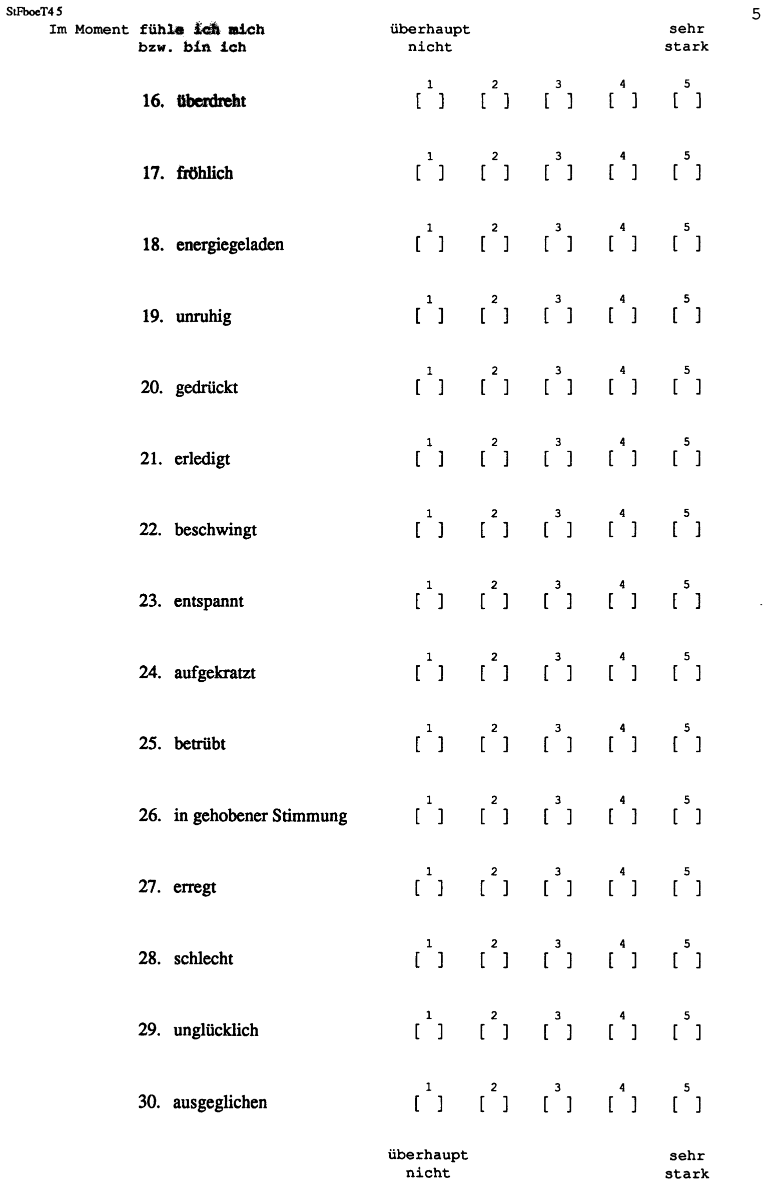
3/3

B.2 MDBF



StFboeT44 4

Im Moment fühle ich mich bzw. bin ich	überhaupt nicht				sehr stark
1. ausgeruht	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
2. zufrieden	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
3. ruhelos	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
4. nervös	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
5. heiter	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
6. schwungvoll	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
7. träge	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
8. ärgerlich	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
9. vergnügt	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
10. angespannt	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
11. gereizt	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
12. erschöpft	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
13. mißmutig	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
14. gut	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
15. reizbar	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
	überhaupt nicht				sehr stark



StFbocT4 5

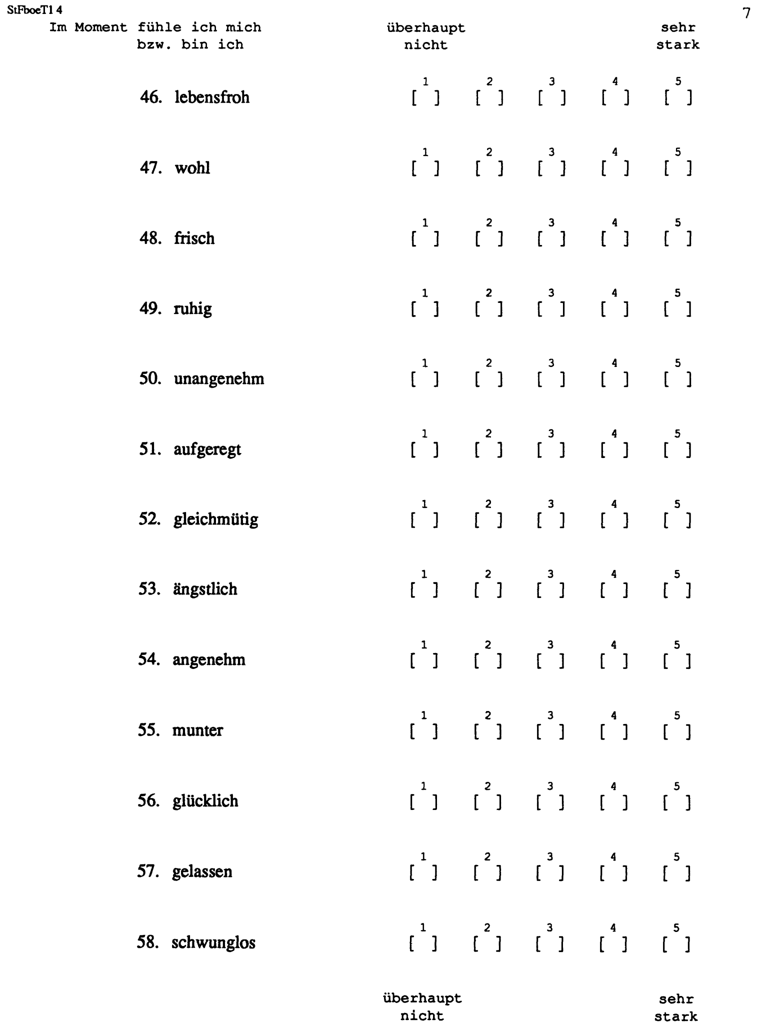
5

Im Moment fühle ich mich bzw. bin ich	überhaupt nicht				sehr stark
16. überdreht	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
17. fröhlich	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
18. energiegeladen	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
19. unruhig	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
20. gedrückt	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
21. erledigt	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
22. beschwingt	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
23. entspannt	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
24. aufgekratzt	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
25. betrübt	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
26. in gehobener Stimmung	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
27. erregt	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
28. schlecht	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
29. unglücklich	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
30. ausgeglichen	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
	überhaupt nicht				sehr stark

StFboeT4 6

6

Im Moment fühle ich mich bzw. bin ich	überhaupt nicht				sehr stark
31. müde	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
32. unwohl	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
33. erregbar	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
34. in gedrückter Stimmung	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
35. blendend	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
36. mißgestimmt	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
37. glückselig	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
38. ermattet	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
39. unzufrieden	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
40. schläfrig	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
41. schlapp	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
42. wach	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
43. ausgewogen	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
44. übermütig	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
45. trübsinnig	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
	überhaupt nicht				sehr stark



StFboeT1 4 7

Im Moment fühle ich mich bzw. bin ich	überhaupt nicht				sehr stark
46. lebensfroh	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
47. wohl	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
48. frisch	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
49. ruhig	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
50. unangenehm	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
51. aufgeregt	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
52. gleichmütig	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
53. ängstlich	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
54. angenehm	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
55. munter	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
56. glücklich	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
57. gelassen	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
58. schwunglos	1 [ ]	2 [ ]	3 [ ]	4 [ ]	5 [ ]
	überhaupt nicht				sehr stark

APPENDIX C

R-Syntax

C.1 Longitudinal Item Selection

```
> #################################
> ####   Stuart Application    ####
> #### Item Selection for MDBF ####
> #################################
>
> #### Preparation ----
> library(stuart)
> library(foreign)
> ####
>
> #### Data Handling ----
>
> # Comprenhensive data description:
> #   http://www.metheval.uni-jena.de/materialien/ges7/ZwischenberichtStufe1.php
> mdbf <- read.spss('http://www.metheval.uni-jena.de/materialien/ges7/ges7.sav',
+   to.data.frame=TRUE, use.value.labels=FALSE)
> demo <- mdbf[,2:3]
> names(demo) <- c('sex','age')
> # Import Item Allocation
> load('labels.rda')
> # Select only MDBF ('beschreibungges7.pdf')
> mdbf <- mdbf[,grep('ST[0-5]',names(mdbf))]
> mdbf <- mdbf[,!grepl('59',names(mdbf))]
> mdbf <- mdbf[,!grepl('U',names(mdbf))]
> #  Recoding
> mdbf[,names(mdbf)%in%unlist(labels[labels$Valence=='negative',4:7])] <-
+   (mdbf[,names(mdbf)%in%unlist(labels[labels$Valence=='negative',4:7])]-6)*-1
> # Save the file
> mdbf <- data.frame(demo,mdbf)
> save(mdbf,labels,file='mdbf.rda')
> ####
>
> #### Facet Definitions ----
>
> # Set Up Factor Structure
> fs <- list(wach1=labels$name[labels$Dimension=='wach'],
+   wach2=labels$name2[labels$Dimension=='wach'],
+   wach3=labels$name3[labels$Dimension=='wach'],
+   gehoben1=labels$name[labels$Dimension=='gehoben'],
+   gehoben2=labels$name2[labels$Dimension=='gehoben'],
+   gehoben3=labels$name3[labels$Dimension=='gehoben'],
+   ruhe1=labels$name[labels$Dimension=='ruhe'],
+   ruhe2=labels$name2[labels$Dimension=='ruhe'],
+   ruhe3=labels$name3[labels$Dimension=='ruhe'])
> # Repeated Measures
> repe <- list(wach=c('wach1','wach2','wach3'),
+   gehoben=c('gehoben1','gehoben2','gehoben3'),
+   ruhe=c('ruhe1','ruhe2','ruhe3'))
> ####
>
> #### Heuristics ----
>
> # Generate empty heuristics table
> heu <- heuristics(mdbf,fs,4,repeated.measures=repe,deposit.on='arcs')
> # Assign heuristic values
> for (n in seq_along(heu)) {
+   for (i in 1:ncol(heu[[n]])) {
+     for (j in 1:nrow(heu[[n]])) {
+       heu[[n]][i,j] <- as.numeric(labels$Valence[labels$name==rownames(heu[[n]])[i]]!=
+         labels$Valence[labels$name==colnames(heu[[n]])[j]])
+     }
+   }
+ }
> ####
>
> #### Objective function ----
> fit <- function(chisq,df,pvalue,rmsea,srmr,crel,rel,cfi,tli) {
+   1 / (1 + exp(-10 * (mean(rel) - .8))) +
+     (.5 - (.5 / (1 + exp(-100 * (rmsea - .05))))) +
+     (.5 - (.5 / (1 + exp(-100 * (srmr - .05)))))
+ }
> ####
>
> #### Run ----
>
> # Compute number of combinations
> combinations(mdbf,fs,4,repeated.measures=repe)
> # Run Item Selection
> sel <- mmas(mdbf, fs, 4,
+   repeated.measures=repe, heuristics=heu,
+   deposit.on='arcs', cores=4,
+   colonies=512, ants=32,
+   evaporation=.95, deposit='ib',
+   fitness.func=fit, item.long.invariance='strong',
+   alpha=1)
> ####
```

C.2 Multiple Group Item Selection

```
> ##################################
> ####   Stuart Application    ####
> #### Item Selection for IPIP ####
> ##################################
>
> #### Preparation ----
> library(stuart)
> ####
>
>
> #### Data Handling ----
>
> # Data available at https://osf.io/tbmh5/
> load('ipip.rda')
> load('itemkey.rda')
> ####
>
> ### Country Selection ----
>
> small <- ipip[ipip$country%in%c('Mexico     ','France     ','Malaysia   '),]
> small$country <- droplevels(small$country)
> ####
>
> #### Facet Definitions ----
>
> fs <- list(Op=paste0('I',itemkey$Full.[substr(as.character(itemkey$Key),1,1)=='O']),
+   Co=paste0('I',itemkey$Full.[substr(as.character(itemkey$Key),1,1)=='C']),
+   Ex=paste0('I',itemkey$Full.[substr(as.character(itemkey$Key),1,1)=='E']),
+   Ag=paste0('I',itemkey$Full.[substr(as.character(itemkey$Key),1,1)=='A']),
+   Ne=paste0('I',itemkey$Full.[substr(as.character(itemkey$Key),1,1)=='N']))
> ####
>
>
> #### Heuristics ----
>
> # Generate empty heuristics
> heu <- heuristics(small,fs,2,deposit.on='arcs')
> # Assign Fisher-z correlations
> for (n in seq_along(heu)) {
+   for (i in 1:ncol(heu[[n]])) {
+     for (j in 1:nrow(heu[[n]])) {
+       heu[[n]][i,j] <- as.numeric(itemkey$Facet[itemkey$Full.==substr(rownames(heu[[n]]),2,4)[i]]!=
+       itemkey$Facet[itemkey$Full.==substr(colnames(heu[[n]]),2,4)[j]])
+     }
+   }
+   cors <- cor(small[,colnames(heu[[n]])])
+   cors[abs(cors)<.2] <- 0
+   cors <- abs(.5*log((1+cors)/(1-cors)))
+   heu[[n]] <- heu[[n]]*cors
+ }
> # Remove items without cors > .2
> filt <- lapply(heu,function(x) rowSums(x,na.rm=TRUE)!=0)
> for (i in seq_along(fs)) {
+   heu[[i]] <- heu[[i]][filt[[i]],filt[[i]]]
+   fs[[i]] <- fs[[i]][filt[[i]]]
+ }
> ####
>
> #### Objective function ----
>
> fit <- function(chisq,df,pvalue,rmsea,srmr,cfi,tli,crel,rel) {
+   (1 - (1 / (1 + exp(-100 * (rmsea - .05))))) +
+     (1 - (1 / (1 + exp(-100 * (srmr - .05))))) +
+     (1 / (1 + exp(-25 * (min(unlist(rel)) - .4))))
+ }
> ####
>
> #### Run ----
>
> sel <- mmas(small,fs,2,grouping='country',
+   item.invariance='ess.equivalent',item.group.invariance='strong',
+   heuristics=heu,deposit.on='arcs',
+   fitness.func=fit,,deposit='gb',
+   evaporation=.95,ants=64,colonies=512)
> ####
```

C.3 MTMM Item Selection

```
> #################################
> ####   Stuart Application    ####
> #### Item Selection for EMEX ####
> #################################
>
> #### Preparation ----
> library(stuart)
> library(foreign)
> ####
>
> #### Data Handling ----
>
> load('emo.rda')
> # Item Allocation
> key <- data.frame(item=1:28,word=NA,cate=NA,vale=NA)
> key$word <- c('affection','joy','fear','anger','shame','sadness',
+   'love','happiness','worry','resentment','guilt','depression',
+   'disgust','intimacy','nervousness','cheerfulness','concern','fury',
+   'regret','sorrow','caring','contentment','anxiety','rage',
+   'embarassment','unhappiness','pride','tension')
> tmp <- data.frame(
+   c(7,1,14,21,2,8,16,22,3,9,17,23,4,10,24,18,5,11,19,25,6,12,20,26,13,27,15,28),
+   rep(c('love','joy','fear','anger','shame','sadness',NA),each=4))
> key$cate[tmp[,1]] <- tmp[,2]
> key$cate <- factor(key$cate,labels=levels(tmp[,2]))
> key$vale <- 'neg'
> key$vale[tmp[,1]][c(1:8,26)] <- 'pos'
> ####
>
>
> #### Facet Definitions ----
>
> # Set Up Factor Structure
> fs <- list(posS=paste0('zea',str_pad(key$item[key$vale=='pos'],2,'left','0')),
+   negS=paste0('zea',str_pad(key$item[key$vale=='neg'],2,'left','0')),
+   posA=paste0('zeaa',str_pad(key$item[key$vale=='pos'],2,'left','0')),
+   negA=paste0('zeaa',str_pad(key$item[key$vale=='neg'],2,'left','0')))
> # MTMM Structure
> mtmm <- list(pos=c('posS','posA'),
+   neg=c('negS','negA'))
> ####
>
> #### Objective function ----
> fit <- function(chisq.scaled,df.scaled,pvalue.scaled,rmsea.scaled,srmr,cfi.scaled,tli.scaled,con,crel,rel) {
+   (1 - (1 / (1 + exp(-100 * (rmsea.scaled - .05))))) +
+     (1 - (1 / (1 + exp(-100 * (srmr - .05))))) +
+     (1 / (1 + exp(-25 * (min(unlist(rel)) - .7)))) +
+     (1 / (1 + exp(-10 * (con - .5))))
+ }
> ####
>
>
> #### Run ----
>
> # Compute number of combinations
> combinations(emo,fs,4,mtmm=mtmm)
> # Run Item Selection
> sel <- list()
> for (i in 1:5) {
+   sel[[i]] <- mmas(emo,fs,4,fitness.func=fit,mtmm=mtmm,
+     analysis.options=list(estimator='mlr'),
+     evaporation=.8,ants=16,alpha=1,deposit='ib')
+ }
> do.call(rbind,lapply(sel,function(x) x$log[which.max(x$log$pheromone),]))
> ####
```

Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit allein verfasst und keine anderen, als die angebenen Hilfsmittel verwendet habe. Diese Arbeit ist in keinem früheren Promotionsverfahren angenommen oder abgelehnt worden.

Berlin, 26. Februar 2017

Martin Schultze