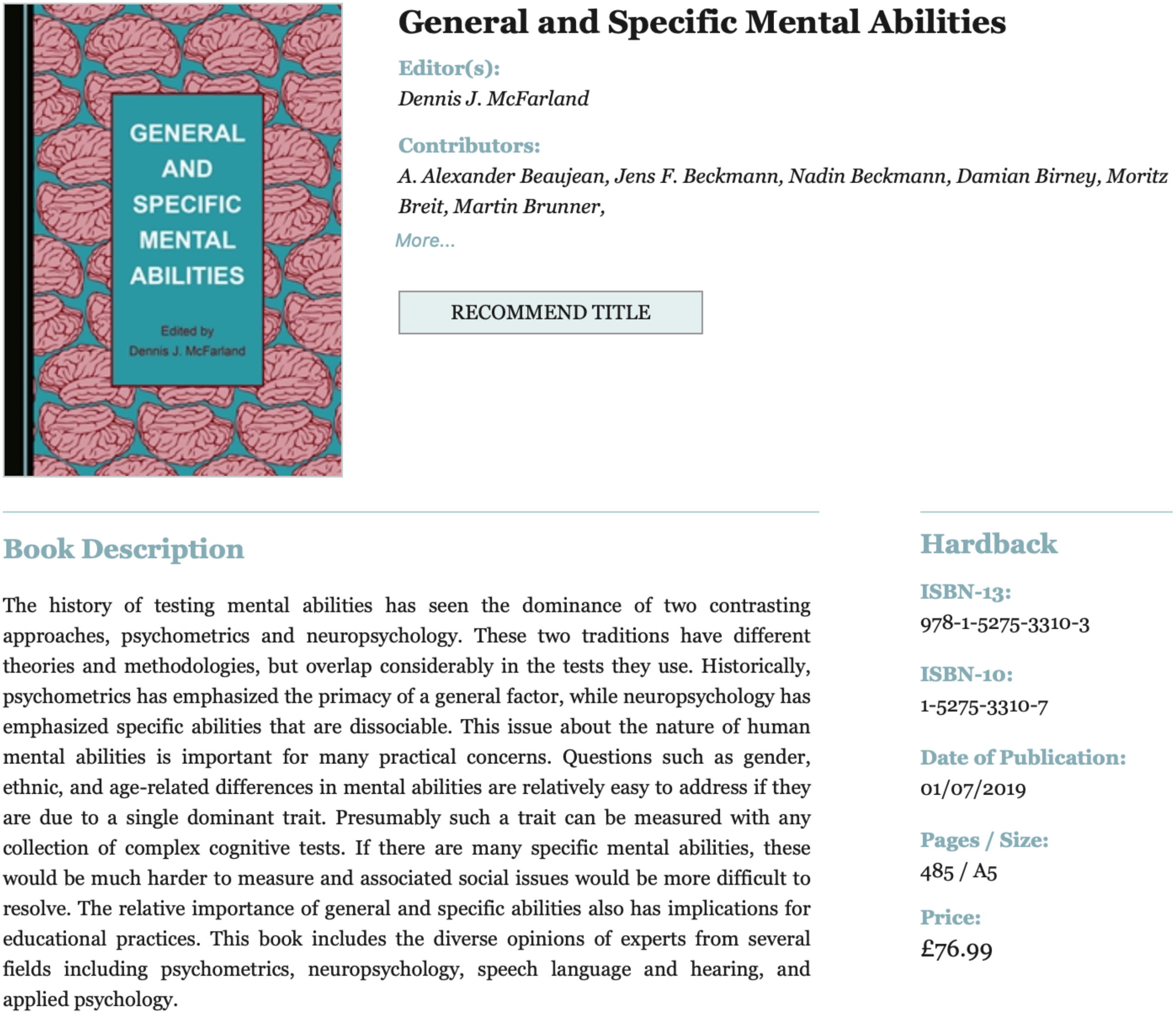

General and Specific Mental Abilities

Editor(s):

Dennis J. McFarland

Contributors:

A. Alexander Beaujean, Jens F. Beckmann, Nadin Beckmann, Damian Birney, Moritz Breit, Martin Brunner,

More...

RECOMMEND TITLE

Book Description

The history of testing mental abilities has seen the dominance of two contrasting approaches, psychometrics and neuropsychology. These two traditions have different theories and methodologies, but overlap considerably in the tests they use. Historically, psychometrics has emphasized the primacy of a general factor, while neuropsychology has emphasized specific abilities that are dissociable. This issue about the nature of human mental abilities is important for many practical concerns. Questions such as gender, ethnic, and age-related differences in mental abilities are relatively easy to address if they are due to a single dominant trait. Presumably such a trait can be measured with any collection of complex cognitive tests. If there are many specific mental abilities, these would be much harder to measure and associated social issues would be more difficult to resolve. The relative importance of general and specific abilities also has implications for educational practices. This book includes the diverse opinions of experts from several fields including psychometrics, neuropsychology, speech language and hearing, and applied psychology.

Hardback

ISBN-13:
978-1-5275-3310-3

ISBN-10:
1-5275-3310-7

Date of Publication:
01/07/2019

Pages / Size:
485 / A5

Price:
£76.99

https://www.cambridgescholars.com/general-and-specific-mental-abilities

General and Specific Mental Abilities

Edited by Dennis J. McFarland

This book first published 2019

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-5275-3310-7
ISBN (13): 978-1-5275-3310-3

CHAPTER TEN

WITHIN-INDIVIDUAL VARIABILITY OF ABILITY AND LEARNING TRAJECTORIES IN COMPLEX PROBLEMS

DAMIAN P. BIRNEY, JENS F. BECKMANN & NADIN BECKMANN

The historical perspective of intelligence is a decidedly between-subjects affair. This is reflected in the dominance of factor analysis as both a psychometric tool for validation and as the cornerstone of the theoretical conceptualisation of intelligence as a hierarchically structured human attribute (Thurston 1938, Horn and Cattell 1966, Carroll 1993, Stankov 2000b, Schneider and McGrew 2012, McGrew 2009). In spite of the significant gains made over the last 120 years in our understanding of its structure, it turns out that knowing what intelligence is and is not correlated with—the psychometric approach to mapping the nomological network (Borsboom, Mellenbergh, and van Heerden 2004, Sternberg 1990)—does not actually tell us much about the basis of intelligence. In this chapter we have a simple objective: to reflect on insights gained in our use of linear mixed-effects models and experimental manipulations to investigate how a *within-subject, process-oriented* approach to human intellect might better augment our understanding of its correlates.

In this chapter we first briefly remind ourselves of the foundations of the psychometric approach underlying the Cattell-Horn-Carroll (CHC) theory of intellectual abilities and how this framework continues to evolve (Schneider and McGrew 2012, Schneider, Mayer, and Newman 2016). We then aim to substantiate why the psychometric approach will always provide a limited account of intelligence and what might be done to redress this. One of the particularly interesting features of intelligence tests is the role of complexity, and its corollary, that intelligence is needed to meet the challenges of complexity in everyday problems. However, what is difficult, is not always complex, so it is important to be clear of the distinction between difficulty and complexity, and we summarise our view on this. Finally, we present three case studies as (1) the basis of an argument for the importance of considering a process-oriented account of the impact complexity manipulations have on performance, and (2) as an example of how this might be achieved using repeated-measures designs and linear-mixed effects regression. We conclude with a description of the core components of *psychometric complexity* as a paradigm for ongoing investigation.

A Hierarchical Perspective on Intelligence: The Psychometric Approach

The CHC theory provides an extensively validated framework for conceptualizing and measuring human intellectual abilities (Schneider and McGrew 2012, McGrew 2009). Its foundation is Spearman's (1904) recognition of the theoretical importance of positive manifold—that all cognitive tasks tend to be more or less positively correlated with each other. Spearman suggested that this correlation reflected a general *mental energy,* or *'g'*. Subsequent research (e.g., Horn and Cattell 1966, Stankov 2000b, Thurston 1938) into a diversity of cognitive tasks demonstrated that performances on some types of tasks tended to be more highly correlated with each other, than they were with performances on other types of tasks. Careful analysis of these statistically 'similar and different' tasks gave insight into potentially common and distinct functions, in addition to (or instead of) 'g' (Carroll 1993). The observed patterns of convergent and divergent correlations were directly interpreted as the manifestation of distinct, fundamental, latent cognitive abilities. Over time, these abilities mapped out the nomological network of intellect into a dynamic, three-stratum taxonomical hierarchy, known as the Cattell–Horn–Carroll (CHC) theory of cognitive abilities (Schneider and McGrew 2012). At the third (top level) stratum is 'g'. A small number of 'broad abilities' define the second stratum, and a larger number of 'narrow abilities' occupy the lowest level or first stratum. McGrew (2009) considered the hierarchy "dynamic" not because the nature of the functions change in degree or type, but because new narrow and broad ability factors can be added to the taxonomy conditional on them meeting this validation standard across multiple samples and contexts.

An Argument for Process-Oriented Accounts

Notwithstanding the extraordinary success of CHC theory in describing between-subject differences, it has long been recognised that the individual-differences approach to the investigation of psychological attributes generally, and intellectual abilities specifically as we have just described, is incomplete without a consideration of process-oriented accounts (van der Maas et al. 2017, Cronbach 1957, Deary 2001). Lohman and Ippel (1993, p 41) citing Cronbach (1957), McNemar (1964), Spearman (1927) and others, concluded that a major reason why the individual differences approach to the study of intelligence

> "…was unable to achieve one of its central goals: the identification of mental processes that underlie intelligent functioning", was because "… a research program dominated by factor analysis of test intercorrelations was incapable of producing an *explanatory* theory of human intelligence".

They argued for a considered cognitive approach where tasks are designed to detect theoretically specified, qualitative differences (see also, Deary 2001). Lohman and Ippel (1993, p 42) were suggesting that the general idea of test theory as applied statistics (i.e., psychometrics) not only hampered the development of *structural theories* for the measurement of processes, but actually precluded it. This was consistent with their reading of the earlier recommendation Guttman (1971) had proposed in his presidential address to the annual meeting of the Psychometric Society. Here, Guttman contrasted the purpose of observation in the psychometric testing tradition, which was (and generally still is) to compare individuals, with his proposed, amended purpose to assess the structure of relationships *among observations*. In effect, Guttman was arguing that if one wishes to better understand the processes of intelligence, one needs to take a distinctively within-subjects perspective. It is precisely this agenda that we explore in this chapter.

There have been many theoretical and technical developments over the last 25 years in particular that have made it easier to address the role of within-subject variability, we will consider some shortly. Yet, the breadth and impact of what psychometric tests of *between-subject intellect* predict is truly impressive and hard to ignore (Gottfredson, 2018) – the psychometric tradition has served us well. This ubiquity of prediction is in no small way responsible for the status of intelligence testing at the very top of the historical successes of the psychological testing movement of the 20th century (Schmidt and Hunter 1998). We are certainly not advocating for a discontinuation of the psychometric tradition. Yet psychometric tests do not sufficiently explain *why* or *how* a prediction should hold in the first place. Again, this limitation is well known. Borsboom, Mellenbergh, and van Heerden (2003) provide compelling argumentation that within-subject level processing must be explicitly incorporated in measurement models if we are to substantively link between-subject models of intellect with what is happening at the level of the individual. It is interesting to note that whereas it is generally well-accepted to take a dynamic, situation-dependent perspective on other individual differences attributes, like personality (Mischel and Shoda 1995, Minbashian, Wood, and Beckmann 2010, Wood et al. 2019), this is generally not the case for intelligence. This is likely due to the belief that intelligence tests assess maximal performance (Neisser et al. 1996), with its ensuing assumption that measures of maximal intelligence and their use imply "the existence of a stable or permanent capability" (Goff and Ackerman 1992, p 538). To elaborate on why this is a limited perspective, we reflect briefly on these aspects of the standard psychometric approach to developing a test, because this stability is *ostensibly* antithetical to the notion of within-subject variability.

The Stability Assumption of Intelligence

So why is intelligence commonly thought to be stable, and why might this be a problem? First, to be clear, we are not concerned here with the fact that normative population-based scaling reflects an appearance of stability over time. Similarly, we are not overly concerned with the arguments of Cattell (1987) and others (e.g., Ackerman 2017, McArdle et al. 2002) who suggest that the apparent stability of intelligence is a necessary outcome of aggregating across multiple abilities that have different developmental trajectories. In terms of within-subjects variability, it does not matter too much which level of aggregation one chooses, 'g', broad or narrow. While aggregation may obscure differences, or at worst preclude their consideration, because these effects are observed at the between-subject level, a within-subject perspective of intelligence is precluded either way (Borsboom, Mellenbergh, and van Heerden 2004, Borsboom 2015).

We believe the more important reason why it has been challenging to integrate an inherent within-individual mutability into the conceptualization of intelligence, is because of the limitations in traditional test development methods and the rigidness of tenets that have evolved to service the principles of best-test design (e.g., Pedhazuer and Schmelkin 1991, Wright and Stone 1979). To demonstrate, consider the notion of learning, which has at its core a within-subject conceptualization. It is generally accepted across various domains of education and psychology, that knowledge and expertise is acquired (at least in part) through the motivated (self-regulated) investment of cognitive resources - that is, as a direct product of learning (e.g., Ackerman 1996, Ericsson 2003, Ackerman and Beier 2005). However, the facilitating cognitive abilities (e.g., *Gf*) underlying knowledge acquisition have typically been assumed to be largely immune (or resistant) to training/learning - that is, to exhibit stability. This in spite of tasks, like the Raven's Advanced Progressive Matrices (APM), requiring induction of rules (i.e., learning) on earlier items to best support the induction and application of different rules on later items as a central explanatory process underlying solution (Carpenter, Just, and Shell 1990, Bui and Birney 2014). Technically, a distinct capacity to learn, separate from *Gf*, is not a threat to the stability assumption because this additional capacity would slot in as a new factor in the CHC framework.

Why the Assumption of Stability is Restrictive

As we have alluded to, the stability assumption has historical and somewhat pragmatic origins linked to test design principles. Consider an intelligence test made up of, say, 36 items (like Set II of the APM). Imagine now that there are individual differences in *within-task* learning from item-to-item that exist and operate in ways that *change the nature of the ability being assessed across the test*. In such cases, a non-random source of variance will be added to the measurement. If one considers the typical test-development process, this variance will be reflected in lower reliability estimates because, rather than the test measuring one construct, it will measure at least two reliable but imperfectly correlated ones: (1) individual differences in the primary intellectual ability of interest, and (2) individual differences in a secondary, *within-task* learning factor that might modify in some way the primary ability being measured. If the effect of the latter is strong, then the test will appear unreliable and because reliability is typically considered to be the upper-bound of validity (Pedhazuer and Schmelkin 1991), our confidence in the validity of the test as a whole (as measuring what it purports to measure, Borsboom, Mellenbergh, and van Heerden 2004) will be shaken. In response, the common practice is to screen out items that demonstrate "instability"– that is, to exclude items with lower item-total correlations (or factor loadings), and keep or add items with higher item-total correlations (or factor loadings). Over repeated test-development iterations, the end result is a test that captures a narrowly defined and *static* component of intelligence.

This is not a new problem. The limitations of traditional psychometrics has long been recognized as overly restrictive in areas where assessment of dynamic processes is of interest, for instance, Dynamic Testing (Guthke and Beckmann 2000, Grigorenko and Sternberg 1998), complex-problem solving (Beckmann, Birney, and Goode 2017, Dörner and Funke 2017), and more recently cognitive flexibility (Beckmann, 2014). The point here is that the psychometric principles of best-test design practice are challenged by constructs that are by definition dynamic, fluid and complexly determined by contextual and intra-personal factors. In other words, rather than having stability and item internal consistency as their assessment goal, the central focus is on within-subject variability, or as Guthke and Beckmann (2000, p 22) put it, on "change and lack of homogeneity". The notion of constructs entailing abilities to manage dynamic changes in complexity requires a consideration of what complexity is, to which we now turn.

Complexity as the "Ingredient" of Intelligence

Jensen (1987) has argued that the most undisputed fact about 'g' is that loadings of tasks on this factor are an increasing monotonic function of the tasks' complexity. This has also been observed and reported more broadly by Gottfredson (1997), who noted that the factor analysis of job attributes also produces a corresponding complexity-of-work factor. The basic tenet here is that high g-loadings correspond with performances in tasks, occupations and work that are more complex - broadly defined, complexity is the "active ingredient" in tests of intellect (Gottfredson 2018, 1997, Jensen 1987). Thus the view is that because 'g' entails a capacity to deal with complexity, an independent indicator of complexity are correlations with (or loadings on) measures of intelligence that increase with task complexity but all else being equal, not with increases in difficulty generated by other task features (Spilsbury, Stankov, and Roberts 1990, Stankov 2000a, Birney and Bowman 2009). However, correlations do not provide a clear conception of precisely what it is that makes a task complex (Schweizer 1998). Without a clear theory of complexity, researchers have often been left little option but to either adopt an eclectic approach to defining the cognitive complexity of a task (cf Stankov 2000a), or resort to post-hoc interpretations (Gottfredson 1997). This is appropriate if one's goal is simply to develop tasks that are good-enough measures of intelligence, however, a greater emphasis on process accounts is needed to understand why these tasks “work”. Decomposing complexity seems a good place to start[1].

Difficulty vs Complexity

In the discussion of cognitive abilities and understanding why intelligence tests work, it is useful to make a finer distinction between difficulty and complexity (Beckmann, Birney, and Goode 2017). Difficulty is atheoretical, in that a rank-ordering of test items that are solved by fewer and fewer people tells us little about what make items difficult, just as correlations alone tell us little about complexity. *Difficulty* is a statistical concept captured by indices such as the proportion of people who answer an intelligence test item correctly. It is closely related to traditional concepts of *ability*, in that ability is conversely a function of the proportion of intelligence-test items a person answers correctly, and is thus a “quantifiable level of a person’s success” (Beckmann, Birney, and Goode 2017, p1). *Complexity* on the other hand, is “conceptualized as a quality that is determined by the cognitive demands that the characteristics of the task and the situation impose” (p1). In the next section we consider an extension of this notion, as proposed by Birney and Bowman (2009) and Birney et al. (2017), and consider the concept of *psychometric complexity* to differentiate empirical difficulty effects from more process-oriented accounts of task complexity.

We present three case studies that entail investigations of different complexity manipulations that are either observed or designed with the objective to broaden our understanding of within-subject accounts of cognitive abilities. Case I tests for complexity (vs difficulty) in four different tasks that have different within-task complexity manipulations. Case II considers item-level responses to investigate evidence of complexity in the correlates of the within-subject performance trajectories of item-difficulty and item-order on the APM. Finally, Case III considers a complex-problem solving (CPS) scenario requiring dynamic exploration and decision making to progress an outcome toward some more or less specific goal. Again we investigate evidence of complexity in the

[1] There are limits to the ubiquity of the complexity account. There are certainly tasks that are neither difficult nor complex yet predictive of fluid intelligence. For instance, performance on the well-known, simple perceptual *inspection time* tasks (Deary 2001), or the *finding squares* task (Oberauer et al. 2003), appear to impose minimal storage or processing load, yet are good predictor of *Gf* (Oberauer et al. 2008, Chuderski 2014).

correlates of the within-subject trajectories across explicit, theoretically specified task manipulations and learning opportunities.

Case I: A Within-Subjects Approach to Complexity

Birney and Bowman (2009) aimed to differentiate process-oriented, theory-linked *complexity* factors from other factors that make solution difficult but do not necessarily place higher demands on *Gf*. They investigated *Gf* processes by experimentally manipulating cognitive demands in four reasoning tasks (see Fig. 10-1). Two tasks came from the work of Stankov's individual differences research on the ingredients of *complexity* in fluid intelligence by considering working memory place keepers (WMP, Stankov 2000a, Stankov and Crawford 1993)—a) the Letter Swaps task in which complexity was manipulated in terms of the number of serial, mental permutations required of three letters; and b) the Triplet Numbers task, where complexity manipulations entailed increasing the nature of conjunctive and disjunctive statements in rule validation of number size. The other two tasks were based on an explicit cognitive theory of relational complexity (RC) (Halford, Wilson, and Phillips 1998)—c) the Latin Square task in which relational complexity was manipulated in terms of the RC demand imposed by the requirement to integrate elements of an incomplete 4×4 matrix, and independently, the number of interim solutions to be held in mind (WMP) while doing so (Birney, Halford, and Andrews 2006, Birney et al. 2012), and d) the Sentence Comprehension task in which the degree of centre-embeddedness (RC) was manipulated (Andrews, Birney, and Halford 2006). Two indicators of cognitive demand were considered. The first was the difficulty effect—task solution was expected to become more difficult as complexity increased. The second indicator was the complexity effect described previously. That is, the expectation was that increases in cognitive load would demand concomitantly increased investment of *Gf* resources (Stankov 2000a). This would be evident in a statistically significant monotonic increase in the strength of the association between Gf and task complexity on performance. That is, as complexity increased, the performance of low vs high *Gf* individuals would diverge. This moderation of complexity on the relationship between task performance and Gf we refer to as *psychometric complexity*. This is to make it clear that the foundation of the distinction between difficulty and complexity is that the latter is a testable theoretical statement, whereas the former is an atheoretical, statistical observation.

Replicating the methodology of Stankov and Crawford (1993), repeated-measures analysis of covariance were conducted with Gf (as measured by Raven's APM) as the covariate. The *difficulty effect* was evaluated by testing the main effects of the complexity level manipulation on performance. The *test of complexity* was the linear contrast of the complexity level × Gf interaction effect, which, if statistically significant, was interpreted to be indicative of a monotonic (linear) increasing association with Gf across the ordered levels of complexity. A summary of our reported results are presented in Table 10-1.

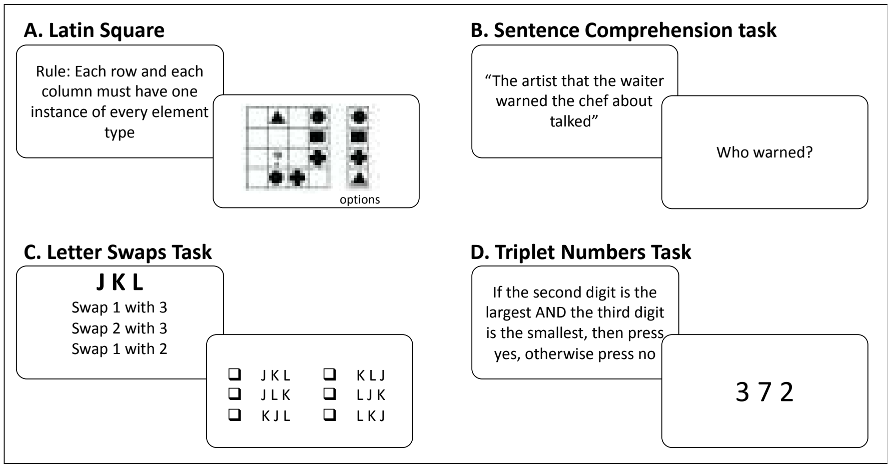
A. Latin Square
Rule: Each row and each column must have one instance of every element type
options
B. Sentence Comprehension task
"The artist that the waiter warned the chef about talked"
Who warned?
C. Letter Swaps Task
J K L
Swap 1 with 3
Swap 2 with 3
Swap 1 with 2
❑ J K L
❑ J L K
❑ K J L
❑ K L J
❑ L J K
❑ L K J
D. Triplet Numbers Task
If the second digit is the largest AND the third digit is the smallest, then press yes, otherwise press no
3 7 2

Fig. 10-1. Example items from Birney and Bowman (2009)

Table 10-1. Summary of partial η^2 effect sizes from ANCOVA reported in Birney and Bowman (2009)

Task	APM[1]	Difficulty[2]	Complexity[3]
Latin Square Task - RC	**.30**[a]	**.68**	.02
Latin Square Task - WMP		**.53**	**.09**
Sentence comprehension test	**.26**	**.35**	**.05**
Letter Swaps Test	**.22**	**.35**	**.12**
Triplet Numbers Test	**.21**	**.38**	**.06**

Notes: 1 = Between-subjects main-effect for APM; 2 = Main-effect for task-level manipulation; 3 = linear contrast of APM x task-level interaction; a = in the LST, RC and WMP (and their interaction) were included in the one analysis along with APM, thus only one effect size is reported; bold: $p < .05$.

As expected, *Gf* was a significant covariate of performance in all four tasks (.21 ≤ partial η^2 ≤ .30; zero-order correlations range: .14 ≤ *r* ≤ .49). Regardless of the complexity manipulation basis, all the tasks became more difficult as demand was systematically increased (.35 ≤ partial η^2 ≤ .88). However, it was not the case that all the manipulations conformed to the psychometric complexity effect as predicted. The point of departure is particularly illuminating because it occurred in the one task (the LST) in which there were two, independent manipulations of cognitive load - the level of relational integration (RC), and the number of interim-steps (WMP) required to be kept in mind during solution (see Fig. 10-2). The RC manipulation in the LST, controlling for number of steps and their interaction, is a manipulation of relational integration load (zero-order correlations range: .45 ≤ *r* ≤ .49). On the other hand, holding multiple interim steps in mind (greater WMP load) was statistically more highly correlated with *Gf* (mean *r* = .44) than problems that could be solved in one step (mean *r* = .34). These results presented evidence of a psychometric complexity effect for the WMP manipulation but not the RC manipulation.

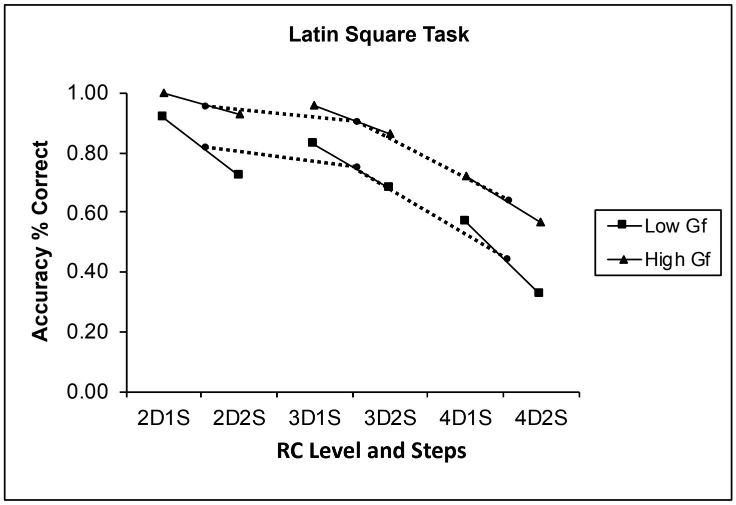

Fig. 10-2. Psychometric complexity with Gf in a Latin Square task. Dashed lines represent RC-manipulation difficulty effect (low- and high-Gf lines are parallel), solid lines represent WMP-manipulation *psychometric complexity* effect (low- and high-Gf lines diverge for each RC level). 2D1S = level 2 RC and 1 step; 2D2S = level 2 RC and 2 step; 3D1S = level 3 RC and 1 step; 3D2S = level 3 RC and 2 step; 4D1S = level 4 RC and 1 step; 4D2S = level 4 RC and 2 step. Adapted from Birney and Bowman (2009).

In subsequent LST research, Bateman, Birney, and Loh (2017) compared a standard WMP load condition with a "dynamic-completion" condition. This latter condition allowed external recording of interim cell solutions that would otherwise need to be held in WM, effectively stripping away critical WMP demand while leaving RC load unchanged. Analyses indicated that performance in the dynamic-completion condition was significantly correlated with *Gf,* even after controlling for standard performance. We see this as converging evidence for Birney and Bowman's (2009) conclusions that WMP manipulations were tapping different aspects of *Gf* from those related to relational processing (i.e., the RC manipulations). What is particularly interesting for our current argumentation, is that Birney and Bowman were able to observe these differential effects using tests of the psychometric complexity hypothesis. That is, while both sets of task manipulations (RC and WMP) impacted performance (performance was better on easier than harder task levels – a *difficulty effect*); and while overall-, RC- and WMP-level specific performance scores were all significantly correlated with *Gf* (a validity test), only WMP task manipulations moderated the *Gf-performance* relationship (a *psychometric complexity effect*).

The LST example in Case I is a part way step toward a more complete within-individual, process-oriented approach to intelligence. It is incomplete because although it considers repeated measures across strong, theoretically underpinned task manipulations, the design is still largely that of the multivariate, psychometric approach. What this work does once again demonstrate however (cf. Lohman and Ippel 1993), is that through systematic task manipulations that represent theoretically specified aspects of cognitive demand (a structural hypothesis, if you will), different parameters from the same task may be isolated for each person. These between-individual differences parameters, derived from within-individual performance differences across task manipulations, can then be submitted to psychometric complexity analyses to investigate and test theories regarding the processes underlying intelligence.

In the next section we introduce an alternative approach based on linear-mixed effects regression (LMER) analyses of within-individual complexity and learning trajectories. We then describe Case II and III which use these methods to derive the parameters needed to test for psychometric complexity effects.

Latent Growth Curve Models and Linear Mixed Effects Regression

Latent growth curve models provide a statistically preferred alternative to within-subjects ANOVA analyses (as used in Case I) and allow more flexibility to consider cognitive correlates in repeated-measures designs. While latent growth curve analyses are typically framed as structural equation models (SEM) (e.g., Schweizer 2006), LMER models, where observations are clustered within individuals, have also been shown to produce more or less equivalent tests (Raudenbush and Bryk 2002, Curran 2003, Hox 2010). Both approaches have the desirable property of modelling growth-curve factors as latent endogenous variables. Whereas SEM approaches provide full flexibility in modelling multiple latent predictors (Curran 2003), LMER models seem to have some advantages in terms of modeling item-linked features in a more straightforward way[2].

In our use of these models, to be described in Case II and III, we are interested in data representing repeated observations (at level 1) clustered within individuals (at level 2). Level 1 data is associated at the level of the item-response, and may include a) the individuals' observed item-accuracy and item-latency, b) item-level aggregates across individuals, such as mean item-difficulty, and c) theoretical specifications like item-complexity (e.g., RC or WMP manipulations) or other item-level "active" ingredient factors. Level 1 data can also include, d) contextual factors associated with the moment of observation, such as the characteristics of preceding items, or item-related metacognitive ratings (e.g., perceived item difficulty, or confidence in the accuracy of one's response), and e) any number of within-level interaction terms. Level 2 data is invariant over level 1 and typically include individual differences factors and between-subject experimental manipulations. The objective of including these variables is to explain (i.e., decompose) variation in intercepts (means) and slopes derived from level 1 variables. We now present two cases where we have used this approach to investigate within-individual performance trajectories in solving complex problems. These trajectories can be defined as differences across the unit metrics of the level 1 variables. In the cases presented we focus particularly on the unit metrics representing item-complexity manipulations and item presentation order. Item-order trajectories are of interest because they can be conceptualised as experience (or learning) curves.

[2] LMER is also potentially more accessible to people outside of the factor-analysis tradition.

Case II: Learning and Ability Trajectories in APM

The aim of the study reported in Case II (Birney et al. 2017) was to model cognitive and non-cognitive correlates of the within-individual trajectories across items of a well-known intelligence test - the APM. We defined two sets of trajectory hypotheses—one according to item-"difficulty" manipulations (to test psychometric complexity) and the other according to item-order (i.e., learning). We then investigated evidence for psychometric complexity in cognitive and personality moderators. As our focus was on repeated measures, it is relevant to reflect on what an individual's item-to-item experience of the APM might look like, and how it fits with our notion of psychometric complexity. To begin, the 36 Set II APM items were designed to progress in cognitive demand according to Raven's (1941) operationalization of intelligence as the capacity required to perceive relations and educe correlates (Spearman 1927)[3]. Like most psychometric tests, item-to-item accuracy accumulates to a total score. In its use, this single score has been demonstrated to be a unidimensional, relatively time-invariant (i.e., stable) indicator of a latent cognitive ability, disconnected from the broader context from which it was collected. In Birney et al. (2017), we argued that from the test-takers perspective, the APM test is an idiosyncratic and very much contextualized experience lasting approximately 40 minutes and likely to be coincident with various dynamic processes that are (assumed to be) "filtered out" in the aggregated total score. In our work, we have been interested in what is happening during that 40 minutes.

Our general goal was to separate the role of learning from performing within APM using multi-level modelling (MLM)[4]. First, controlling for item-to-item *experiences*, we conceptualized *psychometric complexity* as a statistical moderation of the inherent cognitive demand of items according to difficulty trajectories. We used a Rasch calibration to quantify inherent item difficulty, rather than accept the assumption that sequential item-order in an "easy-to-hard" test administration reflects the actual difficulty experienced by participants. Second, controlling for item-to-item difficulty (i.e., the Rasch calibrations), we introduced an additional moderation hypotheses—conceptualized as *psychometric learning*—as the statistical

[3] APM items were ultimately ranked and presented according to the proportion of the standardization sample who answered correctly (i.e., a statistical criteria), rather than by any explicit theoretical hypothesis of the nature of the cognitive resources demanded.

[4] LMER is also commonly referred to as multi-level modelling. We will use this term from here on because it tends to be more descriptive of our use.

moderation of item-order trajectories and analogous to psychometric complexity.

Psychometric complexity hypotheses are tested by the statistical significance of the interaction (i.e., moderation) between cognitive ability (i.e., *Gf*) and the experimental task manipulation indicator. To put this another way, psychometric complexity hypothesises a substantive association between *Gf* and task performance that changes in theoretically meaningful ways depending on the specific level of the task manipulation. Choice of *Gf* as the co-moderator was integral because we specifically designed (or theoretically argued for) task manipulations that demand investment of *Gf* to different degrees. However, the co-moderator need not be a cognitive one. The psychometric complexity paradigm can be applied to any attribute integral to task performance, as long as the demand on this attribute can a) be systematically decomposed into an explicit structural hypothesis, and b) be appropriately implemented and parametrised, ideally via an experimental manipulation. LMER allows for partitioning of performance variability into different sources across different levels of observation. Therefore, psychometric complexity can be tested in non-cognitive components of a cognitive task alongside the cognitive components. In case II, we demonstrate such a decomposition to investigate psychometric complexity effects in both cognitive and personality attributes as co-moderators.

Building a Multi-Level Theory of APM Performance Parameters of Complexity and Their Correlates

Following the nomenclature of Raudenbush, Bryk, and Congdon (2011), an example of the MLM model that Birney et al. (2017) tested is represented in Fig. 10-3 (which we illustrate here using neuroticism as the moderator). Because the outcome is binary (0/1), the model used a logistic link function. There are three parameters of interest at level 1. π_{0i} is the random intercept (mean APM accuracy) for each individual *i*; π_{1i} is the random item-difficulty trajectory (slope) for individual *i*; and similarly, π_{2i} is the fixed item-order trajectory (slope) for individual *i*. Because item-order and item-difficulty are modelled simultaneously, the effect of each is controlled-for by the other (as well as the other variables in the model). There are three moderator-related level 2 parameters central to our main research question. The extent that the moderator variable (in our example, neuroticism) predicts between-individual differences in mean APM is represented by β_{03}. The cross-level interaction between the difficulty trajectory and the moderator variable is β_{12}, and is a parameterisation of the level 2 *psychometric complexity* effect on neuroticism. Finally, β_{22} is the parameterisation of the *psychometric learning* effect on neuroticism because it represents the moderation of the item-order trajectory (changes due to item-order can be conceptualised as learning effects). The basis of the data analysed are the 36-items clustered within each of the *N* = 252 participants (~ 9000+ observations).

Level 1:

$P(Y_{ti} = 1 \mid \pi_i) = \square_{ti}$

$\mathrm{Log}[\square_{ti} / (1 - \square_{ti})] = \eta_{ti}$

$\eta_{ti} = \pi_{0i} + \pi_{1i}.\mathrm{DIFFICULTY}_{ti} + \pi_{2i}.\mathrm{ORDER}_{ti} + e_i$

Level 2:

$\pi_{0i} = \beta_{00} + \beta_{01}.R + \beta_{02}.G + \beta_{03}.N + \beta_{04}.N{\times}G + r_{0i}$

$\pi_{1i} = \beta_{10} + \beta_{11}.G + \beta_{12}.N + \beta_{13}.N{\times}G + r_{1i}$

$\pi_{2i} = \beta_{20} + \beta_{21}.G + \beta_{22}.N + \beta_{23}.N{\times}G$

Where,

DIFFICULTY = Rasch calibrated item difficulty

ORDER = item-order

R = reasoning ability (as a covariate)

G = Group (standard vs confidence)

N = Neuroticism

G *here represents a dummy-coded between-subject experimental manipulation designed as a catalyst for other dynamic processes. Full details can be found in Birney et al. (2017).*

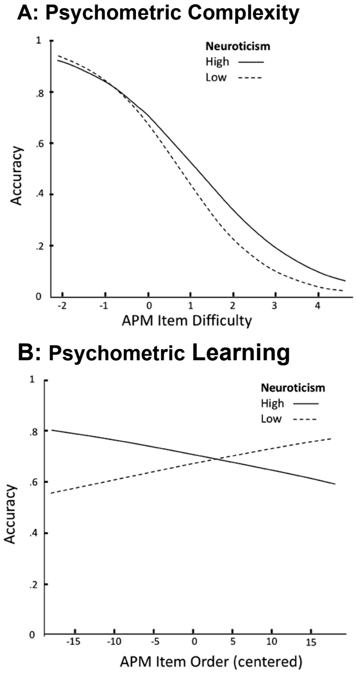

Fig. 10-3. HLM model and representations of the statistically significant cross-level interactions of (A) Item-difficulty × Neuroticism effect (β_{12}) and (B) Item-order × Neuroticism effect (β_{22}). Adapted from Birney et al. (2017, Figure 2).

Findings: As to be expected, reasoning ability (measured independently of APM) accounted for a substantial proportion of the variability in the accuracy of item responses. What was somewhat surprising was that reasoning ability did not moderate the item-difficulty effect (when it was included as the moderator), nor did it differentially predict learning across the task. That is, while those with higher reasoning ability preformed significantly better on APM than those with lower reasoning ability, higher reasoning ability did not proffer any particular advantage in dealing with increased APM item difficulty. This is similar to the Latin Square relational complexity effect reported in Case I, in that there was no psychometric complexity effect for LST RC manipulation either. As for the LST, it may be the case that *reasoning ability* is not what differentiates the *additional* source of *difficulty* in APM from item to item. In the LST, Birney and Bowman (2009) suggested the source of cognitive demand was not relational processing demand, but was instead the capacity for controlled maintenance of information (Kane et al. 2001). In related research, Schweizer and colleagues' (Schweizer 2007, Ren et al. 2014, Ren et al. 2013) report that APM "item-position" effects can be differentially explained by distinct executive functions. Item-position effects are comparable to our item-order parameterisation but derived using SEM approaches rather than LMER models.

As a further brief illustration of the additional information available from MLM analyses, one of the interesting findings of Birney et al. (2017) was that the trajectories of participants' problem-solving were also impacted by individual differences in neuroticism. As represented in Fig. 10-3 (right panel), the analyses suggested that higher (relative to lower) levels of neuroticism proffered an advantage in dealing with increased item difficulty (controlling for item-order), but simultaneously presented a cost to learning as one progresses through the test (controlling for item-difficulty). We suggested these findings were consistent with the dual competing actions account of neuroticism, as simultaneously a propensity for arousal that at medium levels facilitates performance (Szymura 2010, Beckmann et al. 2013), and for anxiety (i.e., worry and test anxiety, e.g., Moutafi, Furnham, and Tsaousis 2006). Further moderation effects of APM trajectories by non-cognitive variables are reported in Birney et al. (2017). We will come back to discuss the implications of these types of moderators in the discussion section. The core point for now is that even in a highly studied, extensively validated task such as the APM, there are substantial insights regarding underlying dynamic processes made more accessible when within-subject analyses are considered that are not available from between-subject total scores.

Case III: Learning and Ability Trajectories in a Microworld

In the last case, we reflect on a dynamic complex-problem solving (CPS) task where we manipulated complexity along different theoretical dimensions, rather than observe it in action (as in Case II). Like complex-problem solving (CPS) tasks in general, the microworlds we used require an active exploration of the problem space to make decisions and observe their impact on outcome variables as the problem-solver aims to reach a more or less specific goal (Wood, Beckmann, and Birney 2009, Dörner and Funke 2017, Funke and Frensch 2007, Greiff et al. 2015). In addition to a cognitive propensity to deal with complexity, CPS tasks require a more dynamic synergy of skills not captured by traditional intelligence tests (like APM), including self-regulation and creativity (Dörner and Funke 2017). In Birney et al. (2018), our core proposition was that if this is the case, then we should see the impact of such conative variables in a CPS microworld when it is framed as a learning task. Again, we took a within-individual, repeated-measures perspective.

The Simulation: The microworld we used was modelled on business stock management processes. The theoretical complexity of decisions was manipulated along two independent dimensions intrinsic to this problem, *delays* and *outflow*. Delays occurred with regard to hiring and firing decisions (framed as being due to time needed to train new hires or due to required notice periods when firing). Outflow of stock, over and above sales (framed as being caused by waste, defects, etc.), was the other variable manipulated. *Delays* have a knowable relational structure. A greater delay between decisions and their impact generates a concomitant increase in cognitive demand, which greater information processing capacity was expected to mitigate. We therefore hypothesised a psychometric complexity effect for delay and reasoning ability on performance. Variable *outflow* (i.e., random around some fixed mean with unknown lower and upper limits) compared to constant outflow, results in less predictable deviations from a targeted stock level. Because of the inherent uncertainty, variable outflow was expected to make the task difficult to manage. However, for the same reasons (i.e., uncertainty), reasoning ability was expected to be less effective in mitigating this type of difficulty (although there may be some strategies that might help, given sufficient motivation to attend to detail). In short, we expected reasoning to show *psychometric complexity* on the delay manipulation, but not on the outflow manipulation.

Eight different variants of the microworld were developed by an incomplete crossing of four levels of delay and three levels of outflow. In all cases, the goal was to reach and maintain a set net inventory level by taking into consideration staffing delays and stock outflow over a period of 30 simulated weeks via the management of the workforce (i.e., number of staff). Each weekly hiring decision constituted a “trial” within the microworld. A “run”, consisting of 30 trials, constituted the 30-week simulated period for a given microworld variant. That is, the multi-level structure was such that multiple trials were nested within runs, and multiple runs were nested within participants. Although this is a three-level structure, the dependant variable was operationalised as the cumulated trial penalty score at the end of a run, which reduced the clustering to a 2-level model (runs and individuals, see Birney et al., 2018, for details). We derived four performance metrics using an analogous MLM approach as described in Case II. Fig. 10-4 represents this model with verbal reasoning as the moderator.

Level 1:

$Y_{ti} = \pi_{0i} + \pi_{1i}.T + \pi_{2i}.R + \pi_{3i}.O + \pi_{4i}.D + \pi_{5i}.O{\times}D + \pi_{6i}.A + e_{ti}$

Level 2:

$\pi_{0i} = \beta_{00} + \beta_{01}.VRT + r_{0i}$

$\pi_{1i} = \beta_{10}$

$\pi_{2i} = \beta_{20}$

$\pi_{3i} = \beta_{30} + \beta_{31}.VRT + r_{3i}$

$\pi_{4i} = \beta_{40} + \beta_{41}.VRT + r_{4i}$

$\pi_{5i} = \beta_{50}$

$\pi_{6i} = \beta_{60} + \beta_{61}.VRT + r_{6i}$

Where,

T = Number of trials completed

R = reasoning ability (as a covariate)

O = Outflow (control vs random)

D = Delay (control vs delay)

A = Attempt number

VRT = Verbal Reasoning Test

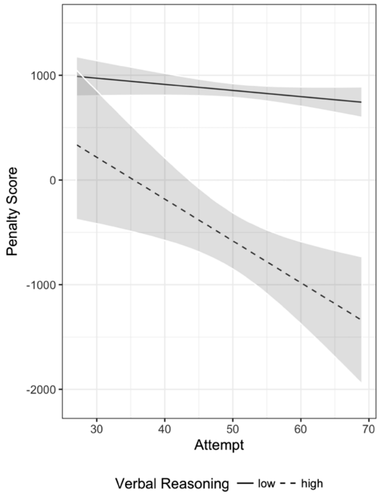

Fig. 10-4. Representation of MLM model applied to investigate within-individual variability in microworld simulation (left) and observed psychometric learning effects on verbal reasoning (right).

The relevant model parameters (as per Fig. 10-4) were a) overall mean performance (*intercept,* π_{0i}), b) competency to successfully manage uncertain outflow complexities (*outflow slope*, $\pi_{3i,}$), c) competency to successfully manage systematic delay complexities (*delay slope*, $\pi_{4i,}$), and d) learning from experience (number of *attempts slope*, $\pi_{6i,}$), with each conditional on the others. A graphical summary of the MLM analyses combining SEM notation is presented in Fig. 10-5 (modified from, Birney et al. 2018, Figure 2). It represents the outcomes of the separate analyses of each moderator of interest (with general reasoning as a covariate in each case).

Analyses were conducted using *R* version 3.4.2 and Linear Mixed Effects (LME) modelling was performed using the *lme4* (Bates et al. 2017) and *lmerTest* (Kuznetsova, Brockhoff, and Christensen 2016) packages. Across the 142 participating mid-level industry managers (i.e., level 2 units), 2116 level 1 observations were available for analysis.

Following the LMER parameter notation detailed in Fig. 10-4, as expected, delays compared to no delays (β_{40}), and variable compared to constant outflow (β_{30}) were associated with significantly higher penalty scores, on average. The effect of delay on penalty scores was significantly more pronounced when outflow was variable (i.e., the within-level interaction of delay and outflow, β_{50}, was statistically significant). Further, performance improved (reduced penalty scores) with increasing number of attempts (β_{60}). The cross-level interactions again represented tests of psychometric complexity and learning. There was evidence for *psychometric complexity* of both general reasoning (when included as the moderator) and verbal reasoning on the delay effect (β_{41}). While higher general reasoning ability did not proffer any benefit *with experience* when included as a moderator[5], specific verbal reasoning ability did (β_{61}, even after controlling for general reasoning ability as a covariate, β_{20}). This moderation is graphed in the right panel of Fig. 10-4. Birney et al. (2017) would refer to this as a *psychometric learning* effect. Within the context of the broader purview afforded in writing this chapter, an alternative proposition would be to consider moderation of learning effects to fall under the general paradigm of psychometric complexity. To achieve this, one would simply consider attempt number as a unitisation of the complexity metric, and proceed to test this as before. We will return to this notion and its implications in the conclusion.

Finally, as expected, in spite of outflow difficulty effects being observed (β_{30}), there was no corresponding *psychometric complexity* of reasoning (narrow or broad) for the outflow manipulation (β_{31}).

In terms of the conative variables, controlling for reasoning ability, the analyses of the personality, motivation, and the emotional regulation variables indicated that they were, by and large, unable to account for any further variation in any of the performance metrics. Interestingly, the few

[5] Using Fig. 10-4 as a reference, the relevant LMER model would replace VRT with general reasoning ability as a moderator. In this separate model, the test of β_{61}, the term representing the moderation of attempt-number (experience) on performance by general reasoning, was not significant.

cases of incremental prediction and psychometric complexity effects of conative variables we did observe were for mindset variables on the outflow manipulation. Performance goal orientation was associated with an unexpected psychometric complexity effect, such that participants with higher performance goal orientations performed better (relative to those with lower performance goal orientations) on the *easier* constant outflow conditions than the more difficult variable outflow conditions[6]. That is, it appeared higher reported performance goals were associated with pronounced performance improvements on aspects of the microworld where capability could be demonstrated and failure avoided. This is theoretically consistent with the extant literature on the distinction between performance- and learning-goal mindsets (VandeWalle 1997, Heslin, Latham, and VandeWalle 2005, Dweck 2000).

The use of a microworld served multiple purposes, but primarily it was chosen to provide a sufficiently dynamic but necessarily structured framework for the investigation of engaged learning and performance under experimentally controlled complexity conditions. Within this context, our findings for domain-general and domain-specific reasoning abilities were largely as expected. General reasoning is important, although the specific nature of the simulation may determine the extent to which investment of domain-specific abilities is beneficial over and above this. Delays, as we define them, fit well with conceptual definitions of working-memory demand (Unsworth and Engle 2007, Birney and Bowman 2009). It seems reasonable to presume that microworld manipulations could be flexibly skewed to target other cognitive abilities. This would allow us to investigate their specific roles in dynamic problem solving also. CPS research tends to address the cognitive aspect of the CPS challenge quite well (Greiff et al. 2015). The broader challenge remains how to incorporate and investigate non-cognitive facets within a complex, dynamic decision making frame (Dörner and Funke 2017). This is particularly true for conative dispositions that have repeatedly been demonstrated to be important to reasoning and learning (Birney et al. 2017, Stankov 1999, Stankov and Lee 2017, Bandura 1997, Zimmerman 2002, Güss, Burger, and Dörner 2017), and where strong claims of incremental prediction for factors such as grit have been made (Duckworth et al. 2007).

[6] This is actually in the reverse direction to the moderation effects we have so far considered, however this does not change its designation as a psychometric complexity effect given it is theoretically coherent (though needs replication).

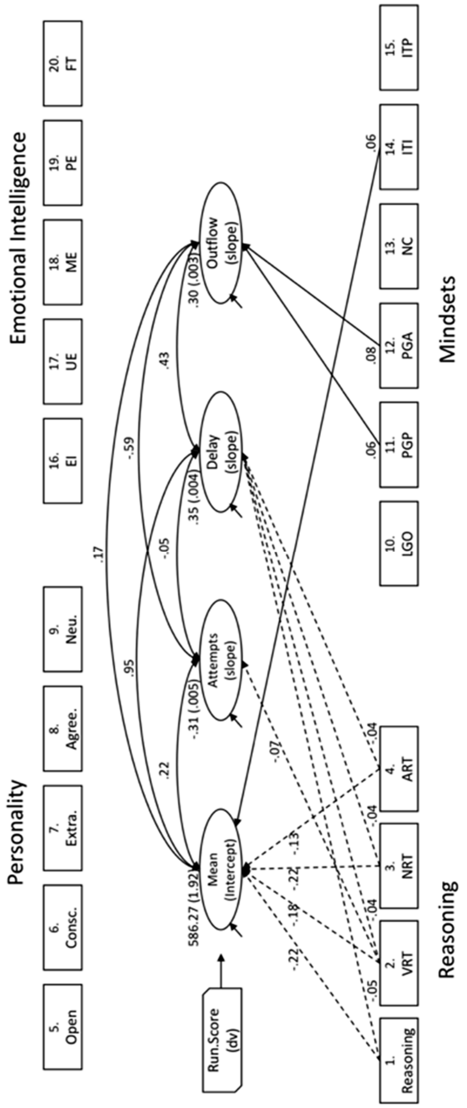

Fig. 10-5. A graphical representation of the Case III MLM model that shows its analogous SEM form. The center represents the LME regression variables. The trial and the delay×outflow interaction term have been omitted because these were not carried through to level 2 (see, Birney et al. 2018, for details).

Why the MLM Approach? A brief note

The MLM approach in Case II and III is distinctly advantageous (Gelman, Hill, and Yajima 2012) compared to the OLS regression/within-subjects ANOVA used in Case I (Brunner and Austin 2009). MLM uses a partial pooling process (often referred to as "shrinkage") that has been shown in simulation studies to shift estimates of the regression coefficients and their associated standard errors toward known mean coefficients (i.e., in simulated population data). This processes has the desirable effect of shrinking coefficients that are estimated with small accuracy more so than those estimated with higher accuracy (Hox 2010), thus intervals for comparisons are more likely to include zero and statistical tests are appropriately more conservative in terms of type 1 errors (Gelman, Hill, and Yajima 2012). MLM models are comparable to the fixed-links SEM models of similar intelligence data discussed by Schweizer and his colleagues (Schweizer et al. 2015, Ren et al. 2012, Wang et al. 2013). Both approaches are consistent with the general class of latent growth-curve models. Curran (2003) and others have demonstrated that MLM analyses are equivalent to latent-growth curve SEM, where the intercepts, slopes and cross-level parameters that come from MLM (when these are allowed to vary as random-effects) are equivalent to the latent-growth variables in SEM (Muthen 1997).

The SEM approach is a suitable and flexible way to analyse the type of data generated by our research, particularly if one were interested in an analysis that included multiple latent-variable moderators at level 2 (because LME regression models are constrained to have a single level 1 dependent variable). However, we argue for staying with the MLM approach for pragmatic reasons. At the time of Curran's writing, it was inordinately clumsy to implement MLM models in the available SEM software. While there have been great improvement since (e.g., MPLUS has been developed to better handle multi-level long data), in our view the LMER approach more intuitively allows for multiple level 1 attributes to be incorporated into the modelling.

Desiderata: Psychometric Complexity as an Investigative Paradigm

The term psychometric complexity has emerged from our analyses of the cognitive abilities literature, including earlier between-individual theories regarding the ingredients of complexity in intelligence (Stankov and Crawford 1993, Gottfredson 2018). In the work summarised here, we have demonstrated psychometric complexity has value in relation to non-cognitive attributes as well — for instance, in Case II for neuroticism and Case III for performance-goal orientation. Because psychometric complexity is linked to structural hypotheses regarding within-individual processes, tests of psychometric complexity have the potential to provide insights into the underlying structure of task performance over and above between-subject investigations.

We hope other researchers might find value in our overall approach to psychometric complexity. Accordingly, in this final section we provide an elaborated specification of the core concepts. First, it seems important that complexity not be confused with complicatedness. Tasks differ in "complicatedness" when, for instance, they a) require different behavioural responses, b) demand different sensitivities to triggers for action, c) prime different motivations and inclinations to act, and d) require different sets of abilities and competencies. Complicated tasks can be difficult for any number of reasons, and as we have argued here and elsewhere (Beckmann, Birney, and Goode 2017, Birney, Beckmann, and Seah 2016), difficulty is an atheoretical, statistical concept. On the other hand, *complexity* is psychologically substantive. It is linked to a single, specific psychological attribute, like *Gf*, but is a *task* quality rather than an aspect of the person. Manipulations along the continuum of task complexity are manipulations of requisite demand on the psychological attribute. In a multidimensional task, we would expect each dimension to have its own continuum of complexity contributing to the overall difficulty experienced by the problem-solver. While other random and unknown systematic factors may also contribute to difficulty, regardless of their number, difficulty has only one continuum—the operationalised task performance continuum. Only in a truly pure, unidimensional task will the complexity continuum coincide with the difficulty continuum. Of course, such tasks do not exist.

Complexity is relative by definition, in that one task has certain complexity relative to another, and one variant of a task (a manipulation) has certain complexity relative to another variant of the same task. To investigate the nature of intelligence, systematic manipulations based on structural hypotheses regarding differential demand on intelligence must be made (Lohman and Ippel 1993). These manipulations are "complexity" manipulations. Therefore, we define psychometric complexity as the extent to which *within-individual* differences in task performance across complexity manipulations differ as a function of *between-individual* differences in that attribute.

Concluding Comments

The core contribution we aimed to make in this chapter was in regard to progressing a broader understanding of between-individual differences (level 2) in within-individual variability (level 1) in complex problem solving. By investigating level 2 correlates, our work is certainly still grounded in the between-subject tradition. However, our focus is on the individual's localised performance trajectories across repeated occasions under different experimental conditions (Lohman and Ippel 1993). We have demonstrated that *within-individual, process-oriented* facets of performance can be identified and studied in novel ways using linear mixed-effects growth-models. Our work and the work of others using related methods (Schweizer 2007, Ren et al. 2014, Ren et al. 2013), suggests that the psychometric complexity paradigm may allow us to better quantify and incorporate more nuanced effects into theory development, and move us one step closer to producing an *explanatory* theory of human intelligence (Lohman and Ippel 1993).

References

Ackerman, P. 1996. "A theory of adult intellectual development: Process, personality, interests, and knowledge." *Intelligence* 22:227-257.

Ackerman, P. 2017. "Adult intelligence: The construct and the criterion problem." *Perspectives on Psychological Science* 12 (6):978-998. doi: 10.1177/1745691617703437.

Ackerman, P., and M.E. Beier. 2005. "Knowledge and intelligence." In *Handbook of Understanding and Measuring Intelligence*, edited by O. Wilhelm and R.W. Engle, 125-139. CA: Sage.

Andrews, G., D. P. Birney, and G.S. Halford. 2006. "Relational processing and working memory in the comprehension of complex relative clause sentences." *Memory & Cognition* 34:1325–1340.

Bandura, A. 1997. *Self-efficacy: The exercise of control*. New York, NY: Freeman.

Bateman, J.E, D. P. Birney, and V Loh. 2017. "Exploring functions of working memory related to fluid intelligence: Coordination and relational integration." In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, edited by G. Gunzelmann, A Howes, T Tenbrink and E.J. Davelaar, 1598-1603. Austin, TX: Cognitive Science Society.

Beckmann, J.F, D. P. Birney, and N Goode. 2017. "Beyond psychometrics: The difference between difficult problem solving and complex problem solving." *Frontiers in Psychology: Cognitive Science* 8:1739. doi: https://doi.org/10.3389/fpsyg.2017.01739.

Beckmann, N, J.F Beckmann, A Minbashian, and D. P. Birney. 2013. "In the heat of the moment: On the effect of state neuroticism on task performance." *Personality & Individual Differences* 54 (3):447-452.

Birney, D. P., J.F Beckmann, N Beckmann, and K.S Double. 2017. "Beyond the intellect: Complexity and learning trajectories in Raven's Progressive Matrices depend on self-regulatory processes and conative dispositions." *Intelligence* 61:63-77. doi: https://doi.org/10.1016/j.intell.2017.01.005.

Birney, D. P., J.F Beckmann, N Beckmann, K.S Double, and K Whittingham. 2018. "Moderators of learning and performance trajectories in microworld simulations: Too soon to give up on intellect!?" *Intelligence* 68:128-140. doi: https://doi.org/10.1016/j.intell.2018.03.008.

Birney, D. P., J.F Beckmann, and Y Seah. 2016. "The eye of the beholder: Creativity ratings depend on task involvement, order and methods of evaluation, and personal characteristics of the evaluator." *Learning and Individual Differences* http://dx.doi.org/10.1016/j.lindif.2015.07.007 (51):400-408.

Birney, D. P., and D. B Bowman. 2009. "An experimental-differential investigation of cognitive complexity." *Psychology Science Quarterly* 51 (4):449-469.

Birney, D. P., D. B Bowman, J.F Beckmann, and Y Seah. 2012. "Assessment of processing capacity: Latin-square task performance in a population of managers." *European Journal of Psychological Assessment* 28 (3):216-226.

Birney, D. P., G.S. Halford, and G. Andrews. 2006. "Measuring the Influence of Relational Complexity on Reasoning: The Development of the Latin Square Task." *Educational and Psychological Measurement* 66 (1):146-171.

Borsboom, D. 2015. "What is cauusal about individual differences? : A comment on Weinberger." *Theory & Psychology* 25 (3):362-368.

Borsboom, D, G.J Mellenbergh, and J van Heerden. 2003. "The theoretical status of latent variables." *Psychological Review* 110 (2):203-219.

Borsboom, D, G.J Mellenbergh, and J van Heerden. 2004. "The concept of validity." *Psychological Review* 111 (4):1061-1071.

Brunner, J, and P.C Austin. 2009. "Inflation of Type I error rates in multiple regression when independent variables are measured with error." *The Canadian Journal of Statistics* 37 (1):33-46.

Bui, M, and D. P. Birney. 2014. "Learning and individual differences in Gf processes and Raven's." *Learning and Individual Differences* 32:104-113.

Carpenter, P. A., M. A. Just, and P. Shell. 1990. "What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test." *Psychological Review* 97:404-431.

Carroll, J. B. 1993. *Human cognitive abilities: A survey of factor-analytic studies*. NY, New York: Cambridge University Press.

Cattell, R. B. 1987. *Intelligence: Its structure, growth and action*. Amsterdam: Elsevier Science Publishers.

Chuderski, A. 2014. "The relational integration task explains fluid reasoning above and beyond other working memory tasks." *Memory & Cognition* 42:448-463.

Cronbach, L.J. 1957. "The two disciplines of scientific psychology." *American Psychologist* 12:671-684.

Curran, P. J. 2003. "Have multilevel models been structural equation models all Along?" *Multivariate Behavioral Research* 38 (4):529-569.

Deary, I.J. 2001. "Human intelligence differences: Towards a combined experimental-differential approach." *Trends in Cognitive Sciences* 5 (4):164-170.

Dörner, D, and J Funke. 2017. "Complex problem solving: What It Is and what It Is not." *Frontiers in Psychology* 8 (1153). doi: 10.3389/fpsyg.2017.01153.

Duckworth, A.L, C Peterson, M.D Matthews, and D.R Kelly. 2007. "Grit: Perseverance and passion for long-term goals " *Journal of Personality and Social Psychology* 92 (6):1087-1101. doi: 10.1037/0022-3514.92.6.1087.

Dweck, C.S. 2000. *Self-theories: Their role in motivation, personality, and development*. Philadelphia, PA: Psychology Press.

Ericsson, K.A. 2003. "The search for general abilities and basic capacities: Theoretical implications from the modifiability and complexity of mechanisms mediating expert performance." In *The psychology of abilities, competencies, and expertise*, edited by R. J. Sternberg and E. L. Grigorenko, 93-125. Cambridge, UK: Cambridge University Press.

Funke, J, and P.A Frensch. 2007. "Complex problem solving: The European perpective — 10 years after." In *Learning to Solve Complex Scientific Problems*, edited by D.H Jonassen, 22-47. NY: Lawrence Erlbaum.

Gelman, A, J Hill, and M Yajima. 2012. "Why we (usually) don't have to worry about multiple comparisons." *Journal of Research on Educational Effectiveness* 5:189-211.

Goff, M, and P. Ackerman. 1992. "Personality-intelligence relations: Assessment of typical intellectual engagement." *Journal of Educational Psychology* 84:537-552.

Gottfredson, L.S. 1997. "Why g matters: The complexity of everyday life." *Intelligence* 24 (1):79-732.

Gottfredson, L.S. 2018. "g theory: How recurring variation in human intelligence and the complexity of everyday tasks create social structure and the democratic dilemma." In *The nature of human intelligence*, edited by R J Sternberg, 130-151. New York: Cambridge University Press.

Greiff, S, M Stadler, P Sonnleitner, C Wolff, and R.B Martin. 2015. "Sometimes less is more: Comparing the validity of complex problem solving measures." *Intelligence* 50:100-113. doi: 10.1016/j.intell.2015.02.007.

Grigorenko, E. L., and R. J. Sternberg. 1998. "Dynamic testing." *Psychological Bulletin* 124 (1):75-111.

Güss, C.D, M.L Burger, and D Dörner. 2017. "The role of motivation in complex problem solving." *Frontiers in Psychology* 8 (851):1-5.

Guthke, J, and J.F Beckmann. 2000. "The learning test concept and its application in practice." In *Dynamic assessment: Prevailing models and applications (Advances in cognition and educational practice)*, edited by C.S. Lidz and J Elliot, 17-69. NY: Elsevier Science.

Guttman, L. 1971. "Measurement as structural theory." *Psychometrika* 36 (4):329-346. doi: http://dx.doi.org/10.1007/BF02291362.

Halford, G.S., W. H. Wilson, and S. Phillips. 1998. "Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology." *Behavioral and Brain Sciences* 21:803-831.

Heslin, P.A, G. P. Latham, and D VandeWalle. 2005. "The effect of implicit person theory on performance appraisals." *Journal of Applied Psychology* 90:842-856.

Horn, J.L., and Raymond B. Cattell. 1966. "Refinement and test of the theory of fluid and crystallized general intelligences." *Journal of Educational Psychology* 57 (5):253-270.

Hox, J.J. 2010. *Multilevel Analysis: Techniques and Applications*. Edited by G.A Marcoulides. 2nd ed, *Quantatitve Methodology Series*. New York: Routledge.

Jensen, A. 1987. "The g beyond factor analysis." In *The Influence of Cognitive Psychology on Testing*, edited by R. R. Ronning, J. A. Glover, J. C. Conoley and J. C. Witt, 87-142. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kane, M.J, M.K Bleckley, A.R. Conway, and R.W. Engle. 2001. "A controlled-attention view of working-memory capacity." *Journal of Experimental Psychology: General* 130 (2):169-183.

Lohman, D. F., and Martin J. Ippel. 1993. "Cognitive diagnosis: From statistically based assessment toward theory-based assessment." In *Test theory for a new generation of tests*, edited by Norman Frederiksen, Robert J. Mislevy and Isaac I. Bejar, 41-70. Hillsdale, NJ: Lawrence Erlbaum Associates.

McArdle, J.J, E Ferrer-Caja, F Hamagami, and R.W Woodcock. 2002. "Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life span." *Developmental Psychology* 38 (1):115-142.

McGrew, K.S. 2009. "CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research." *Intelligence* 37:1-10.

Minbashian, A, R. E. Wood, and N Beckmann. 2010. "Task-contingent conscientiousness as a unit of personality at work." *Journal of Applied Psychology* 9 (5):793-806.

Mischel, W., and Y. Shoda. 1995. "A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure." *Psychological Review* 102 (2):246-268.

Moutafi, J, A. Furnham, and I Tsaousis. 2006. "Is the relationship between intelligence and trait Neuroticism mediated by test anxiety?" *Personality and Individual Differences* 40:587-597.

Muthen, B.O. 1997. "Laten variable modeling of longitudinal and multilevel data." *Sociological Methodology* 27:453-480.

Neisser, U, G Boodo, T.J Bouchard Jr., A.W Boykin, N Brody, S.J. Ceci, D.F Halpern, J.C Loehlin, R Perloff, R. J. Sternberg, and S Urbina. 1996. "Intelligence: Knowns and unknowns." *American Psychologist* 51 (2):77-101.

Oberauer, K., H. Süß, O. Wilhelm, and W. Wittmann. 2003. "The multiple faces of working memory: Storage, processing, supervision, and coordination." *Intelligence* 31:167-193.

Oberauer, K., H. Süß, O. Wilhelm, and W. Wittman. 2008. "Which working memory functions predict intelligence?" *Intelligence* 36:641-652.

Pedhazuer, E.J., and L.P. Schmelkin. 1991. *Measurement, design, and analysis: An integrated approach*. NJ: Lawrence Erlbaum Associates.

Raudenbush, S.W, and A.S Bryk. 2002. *Hierarchical Linear Model: Applications and Data Analysis Methods*. Second ed, *Advanced*

Quantitative Techniques in the Social Sciences Series 1. Thousand Oaks, CA: Sage Publications.

Raven, J C. 1941. "Standardisation of progressive matrices, 1938." *British Journal of Medical Psychology* XIX (1):137-150.

Ren, X., M Altmeyer, S Reiss, and K. Schweizer. 2013. "Process-based account for the effects of perceptual attention and executive attention on fluid intelligence: An integrative approach." *Acta Psychologica* 142:195-202.

Ren, X., F Goldhammer, H Moosbrugger, and K. Schweizer. 2012. "How does attention relate to the ability-specific and position-specific components of reasoning measured by APM?" *Learning and Individual Differences* 22:1-7.

Ren, X., T. Wang, M. Altmeyer, and K. Schweizer. 2014. "A learning-based account of fluid intelligence from the perspective of the position effect." *Learning and Individual Differences* 31:30-35.

Schmidt, F.L, and J.E Hunter. 1998. "The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings." *Psychological Bulletin* 124 (2):262-274.

Schneider, W. J., J. D. Mayer, and D. A. Newman. 2016. "Integrating hot and cool intelligences: Thinking broadly about broad abilities." *Journal of Intelligence* 4 (1):1.

Schneider, W. J., and K. S. McGrew. 2012. "The Cattell-Horn-Carroll model of intelligence." In *Contemporary Intellectual Assessment: Theories, Tests, and Issues*, edited by D Flanagan and P Harrison, 99-144. New York: Guilford.

Schweizer, K. 1998. "Complexity of information processing and the speed-ability relationship." *Journal of General Psychology* 125 (1):89-102.

Schweizer, K. 2006. "The fixed-links model for investigating the effects of general and specific processes on intelligence." *Methodology* 2 (4):149-160.

Schweizer, K. 2007. "Investigating the relationship of working memory tasks and fluid intelligence tests by means of the fixed-links model in considering the impurity problem." *Intelligence* 35:591-604.

Schweizer, K., M. Altmeyer, X. Ren, and M. Schreiner. 2015. "Models for the detection of deviations from the expected processing strategy in completing the Items of cognitive measures." *Multivariate Behavioral Research* 50 (5):544-554.

Spearman, C. 1927. "The abilities of man." In. New York: Macmillan.

Spearman, C. 1904. "General Intelligence," Objectively Determined and Measured." *The American Journal of Psychology* 15 (2):201-292.

Spilsbury, G., L. Stankov, and R. Roberts. 1990. "The effect of a test's difficulty on its correlation with intelligence." *Personality and Individual Differences* 11:1069-1077.

Stankov, L. 1999. "Mining on the "No Man's Land" between intelligence and personality." In *Learning and individual differences: Process, trait, and content determinants*, edited by P. L. Ackerman, P. C. Kyllonen and R D. Roberts, 315-337. Washington, DC: American Psychological Association.

Stankov, L. 2000a. "Complexity, metacognition and fluid intelligence." *Intelligence* 28:121-143.

Stankov, L, and J Lee. 2017. "Self-beliefs: Strong correlates of mathematics achievement and intelligence." *Intelligence* 61:11-16.

Stankov, L. 2000b. "Structural extensions of a hierarchical view on human cognitive abilities." *Learning and Individual Differences* 12:35-51.

Stankov, L., and J. D. Crawford. 1993. "Ingredients of complexity in fluid intelligence." *Learning and Individual Differences* 5 (2):73-111.

Sternberg, R. J. 1990. *Metaphors of mind: Conceptions of the nature of intelligence*. NY: Cambridge University Press.

Szymura, B. 2010. "Individual differences in resource allocation model." In *Handbook of Individual Differences in Cognition: Attention, Memory, and Executive Control*, edited by A. Gruszka, G Matthews and B Szymura, 231-246. New York, NY: Springer.

Thurston, L. L. 1938. "Primary Mental Abilities." In. Chicago, IL: University of Chicago Press.

Unsworth, N, and R.W. Engle. 2007. "The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory." *Psychological Review* 114:104-132.

van der Maas, H, K Kan, M Marsman, and C.E Stevenson. 2017. "Network models for cognitive development and intelligence." *Journal of Intelligence* 5 (16):1-17.

VandeWalle, D. 1997. "Development and validation of a work domain goal orientation instrument." *Educational and Psychological Measurement* 57:995-1015.

Wang, T., X. Ren, M. Altmeyer, and K. Schweizer. 2013. "An account of the relationship between fluid intelligence and complex learning in considering storage capacity and executive attention." *Intelligence* 41:537-545.

Wood, R. E., N Beckmann, D. P. Birney, A Minbashian, J.F Beckmann, and R. Chau. 2019. "Situation-contingent units of personality at work." *Personality & Individual Differences* 136 (1):113-121.

Wood, R.E., J.F Beckmann, and D. P. Birney. 2009. "Simulations, learning and real world capabilities." *Education + Training* 51:491-510.

Wright, B.D, and M. Stone. 1979. *Best test design: Rasch measurement*. Chicago: MESA Press.

Zimmerman, B.J. 2002. "Achieving academic excellence: A self-regulatory perspective." In *The pursuit of excellence through education*, edited by M. Ferrari, 85-110. Mahwah, NJ: Lawrence Erlbaum Associates.