Summer 5-2023

The Examination of Nonparametric Person-Fit Statistics as Appropriate Measures of Response Bias in Ordered Polytomous Items

Mark F. Beck
University of Nebraska-Lincoln, mark.beck449@gmail.com

THE EXAMINATION OF NONPARAMETRIC PERSON-FIT STATISTICS AS APPROPRIATE MEASURES OF RESPONSE BIAS IN ORDERED POLYTOMOUS ITEMS

by

Mark F. Beck

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Psychological Studies in Education

(Quantitative, Qualitative, and Psychometric Methods)

Under the Supervision of Professor Kurt F. Geisinger

Lincoln, Nebraska

May 2023

THE EXAMINATION OF NONPARAMETRIC PERSON-FIT STATISTICS AS APPROPRIATE MEASURES OF RESPONSE BIAS IN POLYTOMOUS ITEMS

Mark F. Beck, Ph.D.

University of Nebraska-Lincoln, 2023

Advisor: Kurt F. Geisinger

Survey research is ubiquitous within the social sciences; however, surveys are vulnerable to response biases. Response biases introduce construct-irrelevant variance into survey responses, which degrades the accuracy of conclusions drawn through the use of surveys. Nonparametric person-fit statistics have been shown to accurately identify response biases in dichotomous response data but are not well studied in polytomous response data. This study examines the accuracy of nonparametric person-fit statistics in polytomous response data. A 6 x 4 x 4 x 2 simulation study was conducted, with type of aberrancy (6), number of response options (4), dimensionality (4), and test length (2) as factors. The sensitivity, specificity, positive predictive value, and negative predictive value for *U3*, the normed number of Guttman errors, and H^T_i were calculated using a bootstrapped cutoff. Findings indicate that these person-fit statistics with a conservative cutoff had excellent specificity but poor sensitivity.

DEDICATION

For Dad -

Mark Douglas Beck

September 4, 1962 – February 7, 2022

TABLE OF CONTENTS

LIST OF MULTIMEDIA OBJECTS

CHAPTER I: INTRODUCTION

Survey research is ubiquitous within the social sciences. Unfortunately, surveys are vulnerable to the introduction of construct irrelevant variance which could lead researchers to draw inaccurate conclusions about a given population. For surveys, response biases are one the most problematic sources of systematic error.

Response biases occur when a respondent systematically answers items in a manner that is independent of the item content. There are six commonly discussed response biases in the extant literature: acquiescence, disacquiescence, midpoint responding, extreme responding, socially desirable responding, and careless responding. While methods exist to identify some of these response biases, traditionally, each response bias requires a unique method. Therefore, using these traditional methods to remove all irrelevant variance due to response biases is impractical. However, research has shown that person-fit statistics, particularly nonparametric person-fit statistics, are useful in identifying response biases in dichotomous data (Dimitrov & Smith, 2006; Emons, 2008; Karabatsos, 2003; Niessen et al., 2016; St-Onge et al., 2011; Tendeiro & Meijer, 2014).

While nonparametric person-fit statistics have been shown to be effective at identifying response biases (i.e., aberrant response patterns) in dichotomous data, there is a dearth of research investigating their use in polytomous data. This lack of research stems from the fact that not many nonparametric person-fit statistics have been generalized for use in polytomous data. Additionally, the use of polytomous data introduces more factors that must be accounted for, for example, different numbers of response categories and multidimensionality. However, nonparametric person-fit

statistics could represent a practical method for identifying response biases in polytomous data.

Given the above, the current study attempted to determine if nonparametric person-fit statistics could be successfully used in polytomous data to identify response biases. To that end, the study simulated polytomous response data with response biases, used nonparametric person-fit statistics (the normed number of Guttman errors, *U3*, and H^T_i) to classify individuals as aberrant or not, and then determined if and when the person-fit statistics were accurate at identifying aberrant responding.

Polytomous response data were simulated using the Multidimensional Graded Response Model (De Ayala, 1994) following recommendations in the extant literature for realistic polytomous data (Bulut & Sünbül, 2017; Jiang et al., 2016). The various response biases were simulated by modifying the item boundary parameters, which is a method employed in similar studies (Emons, 2008; Rossi et al., 2001; Wetzel et al., 2016). Dimensionality, the number of response options, and test length were also included as factors of the simulation.

Once the data were simulated, the nonparametric person-fit statistics were used to classify respondents as aberrant or not. To do so, a bootstrapped, empirical cutoff for each person-fit statistic was found. Using the empirical cutoff, the specificity, sensitivity, positive predictive value, and negative predictive value for each person-fit statistic was found across all unique simulation conditions. Analysis of Variance (ANOVA) was used to determine which factors had a meaningful impact on the accuracy of the person-fit statistics.

Due to the many meaningful three-way interactions, the results were dense and difficult to parse. However, a few patterns still emerged. Please note that all results should be considered within the lens of the meaningful interactions.

One, all person-fit statistics resulted in high specificity and negative predictive value (NPV). This finding suggests that these person-fit statistics, with a bootstrapped cutoff, accurately identified those **not** engaging in aberrant responding. While this is no doubt a result of the conservative cutoff that was chosen, it is still a useful finding. Two, the normed number of Guttman errors showed the best sensitivity and positive predictive value (PPV) overall. However, none of the person-fit statistics showed high sensitivity or PPV with these conservative cutoffs. In terms of the aberrant response patterns, Guttman errors showed the highest sensitivity and PPV when identifying disacquiescence, extreme responding, and careless responding. Coefficient H_i^T showed the highest sensitivity and PPV when identifying midpoint and careless responding. Coefficient *U3* showed the highest sensitivity when identifying disacquiescence, extreme responding, and careless responding.

There were also a few patterns that emerged with regards to the simulation conditions. Again, there were few differences between any of the conditions regarding NPV and specificity. However, in terms of sensitivity and PPV, four and five response options showed the highest outcome estimates, dependent on test length and dimensionality. The medium test length condition almost always showed higher sensitivity and PPV than short test length condition. Dimensionality tended to improve the sensitivity and PPV estimates for five and seven response options as it increased.

Conversely, as dimensionality increased, the sensitivity and PPV for four and six response options tended to decrease.

The findings from this study are comparable to both Emons (2008) and Beck et al. (2019). Emons found that Guttman errors and *U3* performed quite well when identifying extreme and careless responders across test length, the number of response options, and the proportion of aberrant responders in polytomous data. The findings from Emons are supported by the findings from this study, which also found that Guttman errors and *U3* performed well across conditions in polytomous data.

In contrast, Beck et al. (2019) found that H_i^T outperformed Guttman errors and *U3* when predicting a measure of careless responding in real-world polytomous data. Additionally, Beck et al. showed that Guttman errors and *U3* both performed poorly in terms of a ROC analysis and practical impact. This is in stark contrast to the findings of this study, where Guttman errors performed the best at identifying careless responding.

The findings from this study contrast the consensus about the use of person-fit statistics in dichotomous data. Specifically, many studies point to coefficient H_i^T as being the most accurate at identifying aberrant responses in dichotomous data. However, this research suggests that Guttman errors performed better overall (though, this depends on type of response behavior and the characteristics of the survey).

Based on this study, it is difficult to suggest that nonparametric person-fit statistics are accurate indices of aberrant responses in polytomous response data. Using a conservative cutoff (α = .05), the person-fit statistics showed low power when identifying all aberrant response patterns across conditions. However, with a more liberal cutoff, it is likely that these person-fit statistics would show higher power. Regardless, it can be said

that these nonparametric person-fit statistics are accurate indicators of non-aberrant responding in polytomous data. Future research should focus on identifying a standardized cutoff, method for determining a cutoff, or valid null distributions for these nonparametric person-fit statistics. The impact of applying these nonparametric person-fit statistics to real-world data sets should also be examined.

CHAPTER II: REVIEW OF LITERATURE

Survey Research

There are two types of research commonly conducted by educational and psychological scientists: experimental and nonexperimental research. Both types of research have unique strengths and weaknesses and require different interpretations by researchers. Experimental research is defined by control. Namely, control over variables in an experiment and control over the assignment of participants to experimental (or control) groups (Kerlinger & Lee, 2000). The direct control over variables and assignment allows for a very specific interpretation of experimental results: statements of causality. While statements of causality are a strong benefit of experimental research, experimental research has weaknesses. Generally, experiments are more resource intensive than nonexperimental studies, and experiments are not able to address all research questions and/or variables that may be of interest (Kerlinger & Lee, 2000).

Nonexperimental research is defined by lack of control. In contrast to true experiments, nonexperimental researchers often have little or no control over the assignment of participants to groups and are unable to directly manipulate variables of interest (Kerlinger & Lee, 2000). In this way, nonexperimental research is more observational in nature. A researcher can use nonexperimental research designs to investigate relationships between variables of interest, but, at best, can only provide weak evidence of causality through quasi-experiments (Shadish et al., 2002). While it may seem like nonexperimental research has more weaknesses than strengths, it is important to remember that some constructs are best researched through nonexperimental methods. Additionally, nonexperimental research tends to be less resource intensive and less

burdensome to participants than experimental research, which means it can be conducted on a larger scale (Fowler, 2009; Kerlinger & Lee, 2000). Nonexperimental research designs often employ surveys for data collection (particularly in the social sciences; Singleton & Straits, 2009), and research that primarily uses surveys for data collection is often called survey research.

A survey is a method of collecting data on phenomena in the social sciences. Measurement is a basic foundation of science. Without measurement, there would be few opportunities to quantify observations in pursuit of scientific inquiry. While the physical sciences are often able to directly measure the phenomena they are interested in, social scientists are often interested in unobservable, or latent, phenomena (DeVellis, 2012). The field of psychometrics eventually evolved in an effort to measure such latent phenomena accurately and efficiently.

The field of psychometrics focuses on the measurement of psychological and social phenomena. As such, the chief interests of psychometricians are tests, scales, surveys, questionnaires, and/or measures of unobservable social phenomena. Surveys are generally constructed from a set of effect indicators which have answers that are often (but not necessarily) assumed to be theoretically caused by the latent construct of interest (Bollen, 1989; DeVellis, 2012). While developing a survey takes time and resources, once a survey is developed it can easily be leveraged in many types of nonexperimental research designs (Fowler, 2002).

Survey research is ubiquitous within the social sciences; surveys are used in the fields of marketing, psychology, sociology, business, political science, etc. Companies and researchers frequently utilize surveys because they are efficient to build, administer,

and analyze compared to other research methodologies (DeVellis, 2012; Fowler, 2009). Put simply, surveys are one of the most accessible and practical ways to conduct research. However, survey research is not without drawbacks. Data obtained from surveys can be contaminated with much irrelevant information, including error from sampling bias and systematic measurement error (Fowler, 2009; Groves, 1987). Unfortunately, survey data rife with error could lead researchers to draw inaccurate conclusions about a given population. Therefore, it is paramount for researchers who use surveys to identify and, whenever possible, prevent these errors from impacting their results. The sources of error in surveys can be broken down into two general categories: 1) Errors caused by who gives the answers (i.e., sampling bias) and 2) Errors associated with the answers themselves (i.e., measurement error; Fowler, 2009).

Sampling Bias

Sampling bias arises from sampling error, where sampling error refers to irrelevant variance (i.e., both random and systematic variance) introduced into data due to sampling (Shadish et al., 2002). Sampling bias specifically refers to the systematic component of this irrelevant variance. Sampling bias can occur when a nonrandom sample of the population is obtained for a particular study. A nonrandom sample occurs when all members of the population of interest do not have an equal chance to be selected for a given study. Sampling bias can lead to questionable external validity, which is the ability to apply findings from a sample to other samples or the population (Shadish et al., 2002).

It is exceedingly difficult to obtain a true random sample in the social sciences (see Fowler 2009; Kerlinger & Lee, 2000; Shadish et al., 2002, for discussions of why).

However, survey research is often maligned for sampling only a nonrandom fraction of the overall population of interest. In psychology, a large proportion of survey research is conducted on undergraduate students. In marketing, surveys are often only administered to individuals who visit particular websites. In educational psychology, surveys are often administered to easily accessible groups of students. Unfortunately, these only represent a few examples of sampling biases within the social sciences. While these limited samples are often due to practical constraints, it is still deleterious to the external validity of survey research (Shadish et al., 2002). Thankfully, researchers are aware of sampling bias in survey research, and have taken steps to increase the external validity of their studies.

A variety of methods have been employed to address concerns about the generalizability (i.e., external validity) of studies in the social sciences. These studies allow researchers to examine the validity of surveys, and the items that make up surveys, across different socio-cultural categories (i.e., genders, cultures, languages, ages, etc.). Additionally, psychometricians have started to focus on the adaptation of surveys, rather than mere translation, for different cultures and/or languages. While these methods do not directly address sampling bias, they are a step in the right direction. There have also been instances where an institution or government will obtain a representative sample of the population of interest by purposive sampling or true random sampling, though these instances are rare. It should also be noted that scientific replication can be used to address poor sampling techniques; however, with the current state of academia in the social sciences, replication studies are rare (see Bornstein, 1990; Frias-Navarro et al., 2020; Maxwell et al., 2015; Morawski, 2019, for discussions about the “replication crisis”).

Sampling bias is an important issue that affects survey research. However, sampling bias is not a sufficient reason to ignore findings from survey research. As discussed, many survey researchers are aware of these issues and take steps to address concerns about external validity. Additionally, adequately trained researchers are wary of generalizing on the basis of one study. Finally, researchers are trained to be explicitly candid about the limitations of their studies, which includes disclosing any concerns about external validity.

Measurement Error in Classical Test Theory

Measurement error is one of the most discussed and examined threats to the validity of survey research. Indeed, measurement error has been discussed since testing and surveys first became a popular social science research methodology (see Thorndike, 1904, for an early discussion of measurement error). This focus on measurement error is not without good reason. Measurement error can affect the statistical conclusion validity of the surveys being administered (Shadish et al., 2002). Put another way, measurement error can lead researchers to draw poor conclusions from their data. Specifically, measurement error can increase the likelihood of Type I and II errors by increasing the likelihood of over- or underestimating the magnitude of effects, as well as the degree of confidence in the effects (Jaccard & Wan, 1995; Shadish et al., 2002). As such, researchers have made the prevention, identification, and treatment of measurement error a focus within survey research. In fact, measurement error is an important aspect of the theories used by the social sciences to understand surveys and tests.

Classical Test Theory (CTT) posits that a test score is a function of True score and random error (i.e., measurement error). One common representation of this is the true score model, which is given in Equation 1:

$$X_i = T_i + E_i, \quad (1)$$

where X_i is the observed score of person *i* on a set of variables, T_i is, theoretically, the true score for person *i*, or person *i*'s actual level of the construct of interest (mathematically, T_i is the population mean of *Xi*), and E_i is error impacting person *i*'s score (Allen & Yen, 1979; DeVellis, 2012; Fowler, 2009). It should be noted that error and true score are not directly observable. The error can come from many different sources, but it is usually not parsed into further components. Additionally, the error is assumed to be completely random with a mean of zero (Allen & Yen, 1979).

While CTT is an intuitive and useful way to understand measurement, it is not without issues. Measurement under CTT requires certain assumptions be made about data that are easily violated. Specifically, one of the assumptions states that error is not correlated with true score. This assumption is given in Equation 2 below:

$$\rho_{ET} = 0, \quad (2)$$

where, ρ_{ET} represents the relationship between error and true score in a population of examinees (Allen & Yen, 1979). This assumption is violated whenever there is a systematic component to the error term in Equation 1 (E_i). In a sample there are a multitude of ways in which this assumption could be violated: response biases, individual differences (i.e., nonrelevant, between-person characteristics), testing effects (e.g., fatigue), DIF, etc.

CTT also assumes that the measurement errors on two different tests are uncorrelated. This assumption is given in Equation 3:

$$\rho_{E_1E_2} = 0, \tag{3}$$

where, $\rho_{E_1E_2}$ represents the population parameter for the relationship between the error variance of test one and test two (Allen & Yen, 1979). This assumption suggests that there is no systematic variance being diffused across tests. Again, there is a variety of ways in which this assumption could be violated: practice effects, fatigue effects, individual differences, setting effects, method bias, etc. The last major assumption CTT makes about error is that the error variance for one test is uncorrelated with the true score of another test. This assumption is given in Equation 4:

$$\rho_{E_1T_2} = 0, \tag{4}$$

where, $\rho_{E_1T_2}$ represents the relationship between the error and true scores from two test forms in a population of examinees (Allen & Yen, 1979). This assumption is similar to the assumption given in Equation 2. However, it makes the assertion that the error scores on one form are independent of the true scores on another test form. Again, the assumption given in Equation 4 is similar to the assumption in Equation 2 and is violated in many of the same ways. However, it could also be violated if an individual is taking two surveys, and one survey contains information that changes the individual's response(s) on the other survey.

Taken together, it is easy to see why measurement error has received so much attention from researchers. As CTT is one of the most common ways to approach measurement, its vulnerability to assumption violations is a potential barrier to understanding, interpreting, and drawing conclusions based on surveys created under

CTT. Attempts to address this issue can be seen in studies examining DIF, measurement invariance, common method bias, etc. These sources of systematic error need to be identifiable and treatable if CTT is to be considered a useful form of measurement.

Sources of Systematic Measurement Error

There are a multitude of ways that systematic error can be introduced into an observed score in CTT. As such, this section does not provide an exhaustive list of all possible sources of systematic measurement error. Rather, this section focuses discussion on the more prevalent and well-researched sources of systematic error. Whenever possible, these sources of systematic error should be modeled or controlled for when engaging in survey research. For each source of error that is discussed, general methods for addressing the source of error are presented.

Common Method Bias

Common method bias occurs when the method of survey administration causes a difference between the true relationship between the measured constructs and the measured relationship between the constructs (Doty & Glick, 1998). Unsurprisingly, method bias can lead to inflated, or attenuated, relationships between constructs, which can make it difficult to accurately judge convergent and discriminant evidence for validity. Furthermore, it becomes difficult to establish construct evidence for validity (Campbell & Fiske, 1959; Doty & Glick, 1998; Shadish et al., 2002). As constructs are foundational to all areas of science, using, interpreting, and disseminating information about valid constructs should be the highest concern for all researchers (Shadish et al., 2002).

A traditional method that can be used to understand the impact of method bias is the Multi-Trait Multi-Method (MTMM) approach (Campbell & Fiske, 1959). MTMM suggests that construct-related validity evidence should be obtained in two ways: assessing convergent and discriminant evidence (Shadish et al., 2002). Convergent-related validity evidence focuses on trying to establish that the construct of interest shares a high correlation with the same or similar constructs. In contrast, the focus of discriminant-related validity evidence is trying to establish that the construct of interest does not correlate with dissimilar constructs. The MTMM approach also suggests measuring a particular construct with multiple traits and with multiple methods, as the name suggests. The use of multiple traits allows for the examination of convergent- and discriminant-related validity evidence, as discussed above. MTMM prescribes the use of multiple methods as it allows for the parsing of test score variance into more components, one component being variance due to method (Campbell & Fiske, 1959; Doty & Glick, 1998). As a result, test scores under the MTMM lens can be viewed as a function of three sources of variance: random variance (error), trait variance, and methods variance. Put another way, a test score is influenced by random error, the trait being measured, and the way the trait was measured (Campbell & Fiske, 1959; Doty & Glick, 1998).

As evinced by the MTMM approach, method bias can represent a problem to survey research. The systematic error potentially present due to the method of survey administration is a violation of the assumptions of CTT. Indeed, survey researchers will often only use one method of collecting data (i.e., surveys or self-report methods). As a result, survey scores obtained from their studies are less generalizable because they might be confounded by common method bias. While the vulnerability of surveys to method

bias is problematic, it should not be seen as sufficient evidence to condemn the entire methodology (Doty & Glick, 1998). For one, most other methods (e.g., experiments) also have issues with method bias. Though, unfortunate from a validity perspective, most studies and experiments are designed with one indicator, or measure, for a particular variable. Common method bias could be accounted for by improving study design. Specifically, the inclusion of multiple measures of important constructs, as done in the MTMM approach, would alleviate this issue (Campbell & Fiske, 1959; Doty & Glick, 1998; Shadish et al., 2002). Otherwise, researchers should consider the impact that common method bias might be having on their study and temper the discussion of their findings accordingly.

Systematic Error from Item and Questionnaire Design

Another potential source of bias in questionnaire research is poor questionnaire and item design. There is a reason that so many introductory measurement textbooks devote entire chapters to appropriate item design (see Allen & Yen, 1979; DeVellis, 2012; Fowler, 2009, for a few notable examples). Poor item and overall questionnaire design can impact the reliability of the items on a survey and impact the validity of conclusions drawn from a survey (Choi & Pak, 2005; Fowler, 2009; Groves, 1987).

Errors Associated with Item Design. It is often said that good item writing is an art more than it is a science. However, that has not stopped researchers from proposing certain characteristics that are associated with good items (i.e., an item's score shows evidence of being reliable and valid on a given questionnaire). It is often recommended that items are unambiguously worded; descriptive enough to provide adequate detail but are not overly long; written at a 6th grade reading level; only include a single idea within

the same question; are interpreted the same for all respondents; and do not frame or lead the response in any way (Choi & Pak, 2005; DeVellis, 2012; Fowler, 2009). From these few recommendations (the above list is only a few of the posited characteristics of a useful item), we can see that item writing is a difficult task.

The difficulty associated with creating good items is partially responsible for the large amounts of research on the reliability and validity of survey data. For example, studies examining DIF and measurement invariance are investigating the interpretation of items and surveys across socioeconomic and demographic categories. Much research of this type is focused on identifying questionnaires in which these problems exist, and there are a plethora of statistics and methodologies that allow for such investigations. Even outside of survey research, researchers are encouraged (and often required) to discuss the reliability and validity evidence of surveys used in their research.

Errors Associated with Questionnaire Design. While errors associated with questionnaire design are discussed less often in the social sciences than errors associated with item design, it remains an important source of systematic bias in surveys (Groves, 1987; Krosnick & Presser, 2010; Lietz, 2010). For example, there might be an issue with the spacing of responses on a given survey that confuses, or frustrates, respondents. This confusion and/or frustration then impacts the responses they endorse as they take the survey. However, there are still recommendations for how a survey should be formatted and designed. Discussions about survey design range from what response scale is most valid (e.g., Guttman Scaling vs. visual analog scales) to whether or not radio buttons are better than check boxes in online surveys (Couper et al., 2001; DeVellis, 2012). A few of the most widely recommended survey design choices are the survey should be self-

explanatory but still contain instructions; surveys should be as short as possible to reduce the burden on respondents; and surveys should be easy to read and uncluttered (Fowler, 2009).

The study of item location effects has become a popular focus of research on survey design in recent years. Studies examining item location effects look at how the placement of items on a survey might impact respondents. Item location effects are often examined within the context of Item Response Theory (IRT), rather than CTT, but are a concern for surveys developed under both theories. In IRT, item location effects have been found to impact test item difficulties in several contexts (see Meyers et al., 2008; Kingston & Dorans, 1984, for two basic examples). In IRT, accurate item difficulty is important for estimating Theta, or a person's ability on the measured construct. In CTT, the concern is that item location effects might have an impact on the average response. Traditionally, there are two ways to address item location effects: randomizing the order of the items or using multiple questionnaire forms (Schurr & Henricksen, 1983).

While systematic error due to questionnaire design is an undesirable quality for any survey, it is manageable (Krosnick & Presser, 2010; Lietz, 2010). Researchers can investigate surveys they use or create for deficiencies in design and address these deficiencies when possible. Many recommendations for good survey design are common sense (e.g., the survey should be easy to read and uncluttered). However, researchers should also remain aware of the less straightforward threats to survey design (e.g., item location effects).

Systematic Errors Associated with Respondents

A wide variety of errors in survey data can be attributed to respondents, or how the respondents answer items (Fowler, 2009; Groves, 1987). For example, participants may have varying levels of motivation or predilections to answer items in a manner that affects their test or survey scores systematically. Unfortunately, systematic error associated with responses/respondents can be very difficult to control. This type of systematic error is not only difficult to identify, but there is no consensus on the best way to address many of the sources. Some sources of these errors can be attributed to item and respondent interactions, for example, certain groups interpreting items differently than other groups. Other errors can be associated with the survey design and respondent interaction; for example, respondents feeling as if they do not have enough information to answer certain items. These errors can also be attributed to traits of the respondents themselves (e.g., memory). Finally, and perhaps most insidiously, due to some trait, mood, attitude, etc., a respondent can purposefully respond in a way that adds systematic error to their answers (Fowler, 2009); this type of purposeful irrelevant responding is usually referred to as a response bias.

All of these sources of error can lead to problems for survey research. However, some are easier to control than others. For example, errors due to item and respondent interaction, and errors due to survey design and respondent interactions can be modeled using DIF and measurement invariance studies. These methodologies were designed to identify how questionnaires might be different for different groups of respondents. If misinterpretations are occurring for all groups, then the issue lies with survey design and/or item generation. Items that ask participants to recall have often been the target of criticism and discussion, as participants are able to accurately recall some events but not

others (Fowler, 2009). A wide variety of resources providing the best practices for using such items are available to researchers (see Belli et al., 2004; Burton & Blair, 1991; Fowler, 2009, for a few notable examples). Often, the design and implementation of accurate recall items is seen as an item generation issue.

Response biases occur when a participant systematically responds to items independent of the item content (Paulhus, 1991). As mentioned, response biases are, arguably, the most problematic source of systematic error in surveys. For one, there are many different types of response biases, and it is not always clear why or how they originate (few studies have empirically investigated the underlying mechanisms). Therefore, the prevention of response biases through survey, study, or item design is difficult. Additionally, several response biases are innately difficult to identify. For example, careless responding is difficult to detect because it is an umbrella term that subsumes several response styles (Meade & Craig, 2008), and it can be difficult to discern a respondent engaging in midpoint responding from a respondent with a true, neutral level of the construct of interest. Finally, while some response biases are well studied and have had methods developed for their identification (e.g., social desirability), several response biases are not as commonly studied and there is no consensus on the best way to identify them.

Response Biases

As mentioned, some response biases are well studied, and methods exist for preventing or identifying them. However, while methods to address some response biases exist, each response bias tends to require a different solution. The different solution for each response bias often results in researchers only controlling for one or a few response

biases in a given study. Therefore, response biases will continue to threaten and/or weaken conclusions drawn from survey research until more practical methodologies for controlling them are investigated.

For the purposes of this work, the discussion of response biases is limited to self-report surveys. That is, surveys that ask respondents about themselves. However, there are also response biases associated with other-report surveys. That is, surveys in which the respondents answer questions about other people (Wetzel et al., 2016). There are six types of self-report response biases which are commonly identified by researchers: acquiescence, disacquiescence, extreme responding, midpoint responding, socially desirable responding, and careless responding. Recall that response biases cause issues for surveys because they can introduce a source of systematic, construct irrelevant variance into scores. Systematic, construct-irrelevant variance represents a violation of the assumptions of CTT, and this irrelevant variance can attenuate relationships between items and constructs. As a result, the attenuated relationships will degrade the accuracy of conclusions drawn from the impacted scores (Meade & Craig, 2012; Wetzel et al., 2016).

Acquiescence

Acquiescence occurs when respondents tend to respond positively, or in agreement, to items on a questionnaire (Baumgartner & Steenkamp, 2001; Knowles & Condon, 1999; Wetzel et al., 2016). A few theories have been posited for why participants engage in acquiescence responding. For example, acquiescence has been theorized to be either a cognitive or motivational issue (Knowles & Condon, 1999). Couch and Keniston (1960) suggested that acquiescence is a function of respondents' personality traits. Specifically, they suggested that respondents engaging in acquiescence

responding were id-driven, and that they responded impulsively to stimuli because they are constantly seeking out novel and immediate stimulation. As such, these respondents are motivated to respond positively to items to elicit a response.

While many such theories have been offered as an explanation for the motivational component of acquiescence, not many formal examinations of these theories have been made (Knowles & Condon, 1999). When there have been formal inquiries, they generally show relationships between acquiescence and personality scales (Blau & Katerberg, 1982; Ray, 1983). Without empirical research supporting the motivational theory, it is difficult to speculate on the role motivation has in acquiescence responding. However, cognitive theories for acquiescence responding have been the target of more formal inquiry and empirical research.

Knowles and Condon (1999) found that respondents who engaged in acquiescence responding tended to "say yes" more quickly than other types of respondents. However, they also found that increasing the cognitive load on respondents increased the frequency of acquiescence responding. They posited that the relationship between increased cognitive load and acquiescence responding was evidence that acquiescence responding results from a disruption of the Spinozan response process. That is, a Spinozan response process is a two-step process in which an item is comprehended (step one) and then reconsidered (step two). Based on their findings, Knowles and Condon argued that the increased cognitive load was making it difficult for respondents to move past the comprehension step of the Spinozan response process.

While it cannot be definitively stated which theory offers the best explanation for acquiescence, there is more compelling evidence to support the cognitive theory than the

motivation theory. Unfortunately, most response bias research is devoted to identifying aberrant response patterns rather than understanding or preventing them (Knowles & Condon, 1999). While this makes prevention difficult, it has resulted in a variety of methods for identifying acquiescence.

Measuring Acquiescence

The difficulty in measuring acquiescence is finding a way to discriminate between someone who is erroneously responding positively to items and someone who would genuinely have a high score on the construct of interest (Paulus, 1991). An early method that was implemented to identify acquiescence was reverse scoring. In this method, items are chosen at random, and their response poles are reversed. For example, a researcher could take a positively worded, Likert-type item and change the valence. Whereas before, *strongly agree* might have been an endorsement of the construct of interest, now *strongly disagree* is an endorsement of the construct of interest. Theoretically, participants who simply respond in agreement with all items would have nonsensical or balanced scores on the construct of interest (Paulus, 1991). In contrast, individuals who are responding to each item as intended will have accurate scores reflecting their position on the construct of interest.

While reverse scoring seems like an easy method to address acquiescence, it is also not without issue. Specifically, empirical research has shown that reverse scoring items may lead to participants feeling confused or less motivated to respond to the questionnaire in a purposeful manner (van Sonderen et al., 2013). Additionally, empirical research has shown that respondents tend to respond less accurately when items are reversed scored than when all items are worded in a positive manner (Sauro & Lewis,

2011; Schrieshem & Hill, 1981; van Sonderen et al., 2013). For these reasons, contemporary methodology suggests that reverse scoring should not be used to detect acquiescence responding. Instead, contemporary methods, such as structural equation modeling and factor analysis, have been suggested as an alternative, and have been used to measure and study acquiescence effects on surveys (Hinz et al., 2007).

Disacquiescence

Disacquiescence is the polar opposite of acquiescence; disacquiescence occurs when participants tend to respond negatively to items on a questionnaire (Baumgartner & Steenkamp, 2001; Weijters et al., 2013). Despite its similarity to acquiescence, disacquiescence has been the target of less empirical research. The dearth of research is partially because researchers often do not distinguish between acquiescence and disacquiescence. Rather, disacquiescence is often seen as the opposite end of an acquiescence spectrum (Weijters et al., 2013). Interestingly, it has also been suggested that both acquiescence and disacquiescence are more likely to occur when an individual is not sure of the answer (either the correct answer to a question, or when a respondent does not have the self-knowledge necessary to answer a self-report question; Paulus, 1991).

Couch and Keniston (1960) examined both acquiescence and disacquiescence (*yea-saying* and *nay-saying* in their terminology) and their relationship to personality traits. Though the personality theory and measurements Couch and Keniston used are outdated by today's standards, the study still yielded some interesting results. Acquiescence and disacquiescence seemed to be polar opposites on the chosen personality scales. Specifically, disacquiescence was associated with careful

consideration, impulse control, and overall stimulus rejection. At the very least, this association provides some evidence to support the usual supposition that disacquiescence and acquiescence are opposites on a larger spectrum. While little empirical research has investigated the matter, it could be speculated that methods for measuring acquiescence could also be successful at measuring disacquiescence.

Extreme Responding

The extreme response style (ERS) occurs when an individual tends to prefer endorsing extreme responses to items (Greenleaf, 1992; Wetzel et al., 2016). For example, a respondent engaging in ERS will tend to endorse *Strongly Agree* or *Strongly Disagree* for Likert-type items, regardless of their actual level of agreement. Historically, there have been several attempts at measuring and explaining ERS, more so than some other response biases. Some researchers have suggested that ERS is invariant within a person (i.e., someone who engages in ERS will always engage in ERS; Weijters et al., 2010; Wetzel et al., 2013). Other researchers have tried to connect ERS with personality traits (Hamilton, 1968), or country-wide characteristics (e.g., level of development or average IQ; Meisenberg & Williams, 2008). While there is some disagreement in the literature about the causes and correlations of ERS, much of the extant literature agrees that ERS is a methodological problem that needs to be addressed. The issue with ERS is clear: ERS introduces construct irrelevant variance that can affect item means as well as item and construct correlations (De Jong et al., 2008).

Measuring Extreme Response Style

As previously mentioned, ERS has been the focus of much research. As a result, there are multiple methods of measuring ERS that have been used or suggested. Early

attempts at measuring ERS involved estimating proportions of extreme responses. This method essentially treated ERS as an estimated binomial proportion, which allowed a variety of hypothesis tests and descriptive statistics to be examined (Greenleaf, 1992; Gold, 1975). Progression in this technique led to the use of binomial error to obtain a more accurate estimate of a *true* ERS score. That is, the unobservable proportion of items from the item population on which a given individual would endorse an extreme response (Greenleaf, 1992). Along with this method, dedicated ERS measures were created and used. While these were useful, the addition of an extra measure assessing ERS to surveys, or survey batteries, was often criticized as being too expensive and too demanding of respondents (De Jong, et al., 2008).

Criticisms of dedicated ERS scales and the binomial proportion method led to researchers investigating IRT models to measure ERS. IRT models were naturally suited to measuring ERS, as IRT models build in cross-classification of item and person characteristics (De Jong et al., 2008; Jin & Wang, 2014). The potential of IRT to build complex item-person models allows the effect of an individual and an item on a given score to be examined separately. Several different IRT models have been used to identify ERS. For example, the partial credit model (PCM; Moors, 2008), the multidimensional nominal response model (Bolt & Johnson, 2009), an extension of the rating scale model (RSM; Wang et al., 2006), and a generalized PCM model (Jin & Wang, 2014) have all been utilized to address ERS.

Recently, an IRTree approach has been suggested for measuring ERS. IRTree models attempt to model probabilistic outcomes (i.e., responses to an item) as functions of latent decision processes. Put another way, the latent decision processes predict

observed item responses (Böckenholt, 2012, 2017). Unfortunately, while modeling a dichotomous item with one latent process is relatively simple, adding additional decision processes to an item with multiple response options can become complicated rather quickly. Regardless, IRTree models have successfully been used to measure midpoint responding and ERS (Böckenholt, 2017; Jeon & De Boeck, 2019), and appear to be an intriguing new method for measuring response biases. While they are an extremely useful tool for researchers, they may be hard to implement for practical uses on a large scale. IRTree models are a form of latent response/IRT modeling, as the name suggests. These models generally have relatively large sample size requirements (i.e., $N > 1000$; Hambleton, 1989). While obtaining these sample sizes would not be an issue for a large testing or survey program, it may prove difficult for moderate to small scale survey programs and academic researchers.

Midpoint Responding

Midpoint responding (MR) is often discussed in relation to extreme responding (ERS; Böckenholt, 2017; Hernández et al., 2004; Paulhus, 1991; Zhang, 2020). However, rather than preferring to select extreme response options (as in ERS), individuals engaging in midpoint responding tend to endorse item responses in the middle of the response scale (Baumgartner & Steenkamp, 2001; Greenleaf, 1992; Paulus 1991). As MR is often discussed with ERS, it has been the target of much research, either directly or indirectly. This research has been split between investigating the causes and correlates of MR and trying to identify and measure MR. It should also be noted that there is a perennial argument regarding the inclusion of a middle response category (i.e., a response

category that is neither in agreement nor in disagreement with the item stem) on questionnaires in non-cognitive measurement contexts.

Differentiating between MR, ERS, and acquiescence can sometimes be difficult. For example, on a 4-point Likert-type scale, if a respondent selects *Agree* rather than *Strongly Agree* a disproportionate number of times, is this an example of MR or acquiescence? Alternatively, if the respondent selects *Strongly Agree* in the same situation a disproportionate number of times, is it an example of acquiescence or ERS? In both situations van de Vijver and He (2014) suggested that MR and acquiescence are positively correlated. As a result, these response biases can be difficult to parse.

Unfortunately, the causes and correlates of MR have received less attention than other aspects of MR. However, it is still an important avenue of research, and some work has been done. Interestingly, studies investigating the causes and correlates of MR have ranged from examining item readability (Velez & Ashworth, 2007) to aggregate country-level personality traits (He et al., 2014). Regardless of the causes of MR, research has provided compelling evidence suggesting that the middle category of a set of response options does not function the same for all respondents (Hernández et al., 2004). As such, MR is an important response bias to identify when conducting survey research when the response scale contains a middle category.

The most common (and traditional) method of measuring MR is simply looking at the frequency of middle option endorsement for a particular respondent (see He et al., 2014; Velez & Ashworth, 2007, for two notable examples). However, IRT has recently been introduced as a method of modeling MR. In fact, research has shown that the

multidimensional nominal response model and IRTree models are effective when assessing MR in questionnaires (Zhang & Wang, 2020).

Socially Desirable Responding

Socially desirable responding (SDR) is the response style that has received the most attention from researchers, specifically in the realm of personality assessment. Originally, SDR was identified as a nuisance to personality researchers when they were trying to use or develop new scales to measure personality traits (Jackson & Messick, 1958, 1962). SDR occurs when respondents select response options that would reflect positively on them, based on their context and culture (Messick, 1991; Paulus, 1991). For example, a respondent endorsing options related to conscientiousness may select options that indicate higher levels of conscientiousness than that respondent has in reality (assuming conscientiousness is desirable in their particular contexts). Conversely, a participant responding to a substance-abuse questionnaire may select options that reflect lower levels of substance abuse than that participant actually engages in, assuming high levels of substance abuse are seen as undesirable behaviors in their contexts. In effect, a participant's responses may reflect current societal norms and preferences more than the true levels of construct(s) if they are engaging in SDR.

Much research has investigated the potential mechanisms and correlates of SDR. Early investigations of SDR were focused on measurement and identification via separate scales designed specifically to measure SDR (Crowne & Marlowe, 1960; Wiggins, 1964). However, the SDR scales used in these studies often did not correlate with each other, and there was evidence of two factors within these scales even though they were assumed to be unidimensional (Holtgraves, 2004; Paulus, 1984). The two-dimensional nature of

the instruments eventually led to evidence for SDR as a two-factor phenomenon, with the two factors being self-deception and impression management (Messick, 1991; Paulus, 1984). Due to this finding, several theories for the mechanisms underlying SDR were posited; many of these theories were based on Sudman and colleagues (1996) stages of responding.

In addition to theorizing about the correlates and mechanisms of SDR, there have been several attempts at lessening its impact on survey data. Some research has shown that providing respondents with a high degree of anonymity is effective for limiting SDR on questionnaires (Becker, 1976; Paulus, 1991). Jones and Sigall (1971) presented an interesting fake-lie-detector technique, in which respondents were hooked up to a machine that they were told is a "pipeline to the soul". While the fake-lie-detector was moderately successful at curbing SDR, it never became a popular method. There have been several other methods attempting to prevent SDR, but the SDR problem persists in survey data. As such, researchers have also investigated the measurement of SDR so that it can be dealt with *post hoc*.

***Measuring Socially Desirable Responding*.**

The measurement of SDR has traditionally revolved around the creation of scales designed to identify SDR. In fact, a cursory Google Scholar search of social desirability returns a plethora of articles developing and investigating scales for SDR in a variety of areas (see Fischer & Fick, 1993; Jacobson et al., 1977; Kwak et al., 2019, for examples). Of course, some of the most well-known SDR scales are the Balanced Inventory of Desirable Responding (BIDR; Paulus, 1988), Edwards Social Desirability Scale (SD; Edwards, 1957), the Marlowe-Crowne Social Desirability Scale (MCSD; Crowne &

Marlowe, 1960), and the MMPI Lie Scale (L and K; Meehl & Hathaway, 1946). While these scales were developed some time ago, they are still being used and studied by contemporary researchers. Criticisms of these scales suggest their inclusion might increase the burden on participants, but they have been shown to be effective at identifying individuals engaging in SDR.

Careless Responding

Careless responding (CR) is perhaps the most complicated response bias to define. This difficulty is mainly due to CR being used as a catchall category for a variety of response patterns assumed to be the result of similar mechanisms. As such, CR commonly refers to response patterns such as: random responding, low effort or inattentive responding, and uniform responding (Baumgartner & Steenkamp, 2001; Credé, 2010; Huang et al., 2012; Johnson, 2005; Maniaci & Rogge, 2014; Meade & Craig, 2012). However, the underlying aspect that binds these disparate response patterns together is the *content nonresponsivity*: failure to respond to the content of the items, regardless of the resulting response pattern. Content nonresponsivity distinguishes CR from other response biases, as other response biases involve the item content in some manner (Clark et al., 2003; Nichols et al., 1989). For example, respondents engaging in SDR still need to be cognizant of what the item is asking, while a respondent engaging in CR is theoretically completely unaware of the item content.

While the mechanisms of CR are not known, it has been theorized that these response patterns may stem from either the inability to read or correctly interpret items, lack of motivation or ability to respond in a purposeful and thoughtful manner, or having the personality traits of high extroversion and low conscientious (Baumgartner &

Steenkamp, 2001; Hauser & Schwarz, 2016; Meade & Pappalardo, 2013; Nichols et al., 1989).

Researchers have become increasingly focused on CR in recent years as conducting survey research on undergraduate and online samples has become more common. These populations are considered likely to engage in CR because it is assumed that they have a lack of motivation to engage with the content of a survey (Meade & Craig, 2012). Alarmingly, estimates of respondents engaging in CR from these samples are relatively high: ranging from 4% to 73% of respondents. Generally, it is safe to assume that around 10% of a university or online sample will contain patterned responses meeting CR definitions (Beck et al., 2019; Johnson, 2005; Maniaci & Rogge, 2014; Mckibben & Silvia, 2015; Meade & Craig, 2012). This prevalence is equally troubling from a data quality and psychometric perspective. Research has shown that the presence of CR can inflate or attenuate relationships between survey variables, increase measurement error, decrease statistical power, and generally obscure survey data making it harder to draw accurate conclusions (Beck et al., 2019; Credé, 2010; Huang et al., 2014; Maniaci & Rogge, 2014; Mckibben & Silvia, 2015; Meade & Craig, 2012).

Specific Measures of Careless Responding

Since CR represents a large threat to survey research, it is no surprise that researchers have spent considerable effort trying to measure CR. However, one of the characteristics of CR that makes it unique is that several distinct response patterns fall under the CR classification. Necessarily, there are a large number of methods designed to measure CR. While there are some general methods to identify CR, several methods are focused on identifying a specific response pattern associated with CR.

Measuring Uniform Responding. Uniform responding is perhaps the easiest CR response pattern to identify, as it actually follows a discernable pattern. When a participant engages in uniform responding, he or she will select the same response option for large portions of a questionnaire. For example, a participant may select the midpoint option for 50% of the questionnaire or select the *strongly agree* option throughout the whole survey. For this reason, uniform responding is often called *long string responding*, since uniform responding visually appears in data as a set of the same response or category label (i.e., response strings; Johnson, 2005; Meade & Craig, 2012).

Since uniform responding often follows such a recognizable pattern, it is often removed after simple visual inspection during data cleaning, or during analyses when response vectors are analyzed for variance. A respondent who has engaged in uniform responding will generate a response vector with little or no variance in their responses, and these are often automatically removed from psychometric analyses. However, uniform responding is not always so easy to identify. For example, some participants may engage in uniform responding for only 20% of the survey. Other individuals, for example, may select one response option for 30% of the survey, and a different response option for another 30% of the survey. While these data are still contaminated with content irrelevant information, it is much harder to determine through visual inspection or descriptive analyses (Johnson, 2005).

To address this issue, Johnson (2005) developed the *long string index*; a sample-based method of identifying long string cut-off values, above which participants are considered to be engaging in uniform responding. In this method, the frequencies of the longest consecutive strings are examined across the response categories. Similar to scree-

plot analyses, large drops in frequencies are identified as cutoffs, and individuals falling outside of the cutoffs are removed from the data set. While research on the efficacy of the long string index as an overall measure of CR has been mixed, it is often considered effective for screening out respondents who engage in uniform responding (Meade & Craig, 2012).

Measuring Random Responding. Random responding can be parsed into two categories: true random responding and effectively random responding (Credé, 2010). True random responding refers to a response process that involves non-content responding, but also where every response option on a given item has an equal chance of being chosen by the participant. True random responding is a response style devoid of any discernable response pattern, which makes it difficult to identify using traditional data cleaning methods. Due to its nature, true random responding is uncommon in most practical settings (Credé, 2010).

Effectively random responding is more difficult to distinguish from other CR response styles. It is often classified simply as content nonresponsivity, while other researchers have classified it as variations of uniform responding. A classic example of effectively random responding is a respondent that alternates between selecting the first and the last response option throughout a survey. This alternating endorsement is different from uniform responding, as they are not selecting the same response option consecutively, but neither do all response options have an equal chance to be chosen by the participant (Clark et al., 2003; Credé, 2010).

Traditional measures of random responding involve the use of validity scales, response time, or assessing item agreement (Clark et al., 2003; Credé, 2010). Using

validity scales to detect random responding is similar to how specific scales are used to detect SDR. Item agreement involves looking at pairs or sets of items that should have similar responses if a participant was responding purposefully. For example, a questionnaire may contain two items assessing *feeling blue* in respondents. If a participant endorses similar responses to these items, then it can be assumed that they were responding purposefully (Credé, 2010). It should be noted that while random responding has these two specific methods of identification, it is often assessed with general measures of CR.

General Measures of CR

While uniform responding manifests somewhat differently, most other CR response patterns are simply assessed using general measures of CR. Quite a few of these general measures were initially developed to identify specific response patterns (such as item agreement) but were found to be somewhat effective at identifying CR more broadly. These general measures include item agreement or consistency indices, special items, outlier analyses, and more recently, person-fit statistics.

Consistency Indices and Item Agreement. The simplest form of consistency index has already been discussed: two items assessing the same construct are examined for response similarity within a participant. There are also methods that involve looking at correlations between sets of items; many of these methods are based on traditional methods of estimating reliability (e.g., split-half reliability). Unidimensional scales or subscales can be split into even and odd halves and the Spearman-Brown split-half formula can be applied to assess consistency within a respondent. Similarly, synonymous, or antonymous, item indices can be created and assessed for response consistency or

inconsistency (Jackson, 1976; Johnson, 2005; Maniaci, & Rogge, 2014; Meade & Craig, 2012). Consistency indices have been shown to be somewhat effective at identifying random responding and CR more generally, however the sensitivity of these indices has been called into question (Huang et al., 2012; Maniaci & Rogge, 2014).

Special Items. Similar to other response biases, researchers have developed special scales to identify CR. However, rather than taking the form of an entirely new questionnaire (as in SDR), these scales in CR are often composed of a few items. The two most common examples of these special items are *bogus items* and *instructed response items*. Bogus items are designed to essentially have a correct and incorrect answer. Bogus items owe their name to the fact that they often ask respondents ridiculous questions about themselves. For example, consider the bogus item *I died last week* assessed on a 4-point, Likert-type scale. Obviously, to respond to this item correctly respondents should endorse *Strongly Disagree*; if they are responding to the item, they obviously did not die last week. Essentially, *Strongly Disagree* is the correct response to this item. Any endorsement other than *Strongly Disagree* can then attributed to CR. While bogus items make logical sense, in practice they are often interpreted figuratively or affected by context effects, which can lead to participants endorsing an incorrect response more often than would be expected (Meade & Craig, 2012; Schwarz, 1999). For example, a participant may figuratively interpret the item *I died last week* and respond with agreement if he or she had a difficult or stressful week.

Instructed response items are designed with a similar logic to bogus items in a less ambiguous way. This clarity makes them more useful than bogus items. As with bogus items, the goal of instructed response items is to create an item that has a definitive

right and wrong answer. However, rather than making an outlandish statement trying to invoke a specific answer, instructed response items tell the respondent how to respond to the item (e.g., *For this item, please select Strongly Agree*). There is no ambiguity on what constitutes a correct response to this example item, and it is not likely to be interpreted figuratively. While these items can serve as useful indicators of CR, particularly when two or more are included on a survey, they are not recommended to be used as the sole indicator of CR. Additionally, the inclusion of too many instructed response items could serve to frustrate participants, which might result in them engaging in different response biases (Meade & Craig, 2012).

Other Approaches to Measuring CR. There are a variety of other approaches that have been used to identify CR. These approaches include response time, IRT modeling, and outlier analysis. Out of these three remaining methods, response time is the one that has received the most research attention and results are generally positive (Beck et al., 2019; Meade & Craig, 2012; Kong et al., 2007; Soland et al., 2019; Wise & Kong, 2005). However, there is no consensus on how to establish a standardized response time cutoff to differentiate normal and disengaged test-taking behavior. Recommendations range from no recommendation being made, to a vague suggestion of using an empirically derived cut-off (Beck et al., 2019; Huang et al., 2012; Meade & Craig, 2012; Niessen et al., 2016). Of note, a few methods of determining a response-time cutoff for measuring disengagement have been empirically investigated, but a consensus has not been reached (Kong et al., 2007; Soland et al., 2019). While response time is no doubt a useful measure of CR, more research is needed to standardize its use.

IRT modeling is one of the newer ways of identifying CR. Specifically, Jin et al. (2018) proposed using mixture IRT modeling to remove the biasing influence of CR on survey data. According to the authors, mixture IRT modeling combines latent trait models (i.e., traditional IRT) and latent class analysis, which allows for respondents to be separated into different latent classes based on their response patterns. While the proposed model is flexible and could theoretically be used to model several response biases, doing so requires that the probability of endorsing any given response option is specified *a priori*. While this method may be useful in certain situations (e.g., when true random responding is suspected), in many situations the probability of selecting certain response options is unknown in individuals who engage in aberrant responding. However, these models represent an exciting new area of research that could have great potential as more research is done.

Finally, outlier analysis has been used to identify CR. Specifically, Mahalanobis distance (MD) has been used in several studies attempting to identify CR. While the research on the efficacy of MD to detect CR is sparse, research suggests that it is not a useful indicator of CR. Namely, while it is effective at detecting aberrant responses in specific contexts, it is not as effective as other measures of CR (Hong et al., 2020; Meade & Craig, 2012). Additionally, it has been shown that MD does not perform well when there are a large number of Likert-type items to be analyzed (Hong et al., 2020; Meade & Craig, 2012). While more research is needed to confirm whether MD is useful as an indicator of CR, it currently appears that is less effective than other methods.

Thus far, many more methods of detecting CR have been discussed than for any other response bias. This disparity is largely due to the glut of methods that exist for

detecting CR, which is a direct result of the multiple response patterns that can occur under the current definition of CR. However, there is one additional measure of CR that must be discussed: person-fit statistics. Person-fit statistics are an expansive topic, as there are many person-fit statistics that have been investigated for use in detecting CR. More than that, several person-fit statistics have been investigated as *general* measures of response biases and aberrant responding: capable of detecting multiple response biases and aberrant response patterns simultaneously. If true, person-fit statistics have great potential for use in survey research. For these reasons, the discussion on person-fit statistics is relegated to its own section.

Person-fit Statistics

Person-fit statistics are a vast category of indices developed to identify improbable or comparably aberrant response patterns. The ability to examine an individual response set and determine improbability is what makes person-fit statistics such powerful tools for identifying response biases. Whereas traditional methods to identify response biases are generally developed to detect a singular bias, person-fit statistics can simultaneously find response patterns that are improbable (as they would be in cases of random or careless responding), or too probable (as they would be in cases of socially desirable responding, acquiescence, extreme responding, etc.). Person-fit statistics have even been used to identify individuals that may be cheating or guessing on academic tests (Levine & Rubin, 1979).

Generally, a person-fit statistic takes an individual response pattern and compares it to a measurement model or a group of other response patterns. If the response pattern in question deviates from the measurement model, or the group, in a significant way, the

person-fit statistic identifies it as improbable or aberrant (Meijer & Sijtsma, 1995, 2001). For a simple example, consider an individual who completed a 10-item questionnaire with dichotomous yes/no response options. If the questionnaire fits a particular parametric IRT model, this individual's response vector could be classified as likely or unlikely given the model. Alternatively, this individual's response vector could be compared to other individuals from the same population who have also completed the questionnaire. If this individual's response vector is vastly different from the group (e.g., the individual endorsed 8 or 9 *yes* responses while most respondents in the group endorsed 1 or 2 *yes* responses) then they could be classified as having an aberrant response set.

As alluded to, most person-fit statistics are derived from IRT models: either parametric or nonparametric. Parametric person-fit statistics generally involve looking at the likelihood of a given response set based on the underlying IRT model, while nonparametric person-fit statistics often involve comparing a response set to a group of response sets to assess aberrancy (Meijer & Sijtsma, 1995). Both parametric and nonparametric person-fit statistics have been used to identify aberrant response patterns with varying levels of success. While most parametric and non-parametric person-fit statistics were developed, and are traditionally used, with dichotomous response data (i.e., correct/incorrect), all of the person-fit statistics discussed in this dissertation have been generalized for use with polytomous response data.

Parametric Person-Fit Statistics

As discussed, most parametric person-fit statistics tend to rely on an underlying IRT model to assess aberrancy. Any response set that is highly improbable given the

underlying IRT model is considered to be aberrant. This measure of probability can be obtained in one of two ways: examining the residuals between the model expected and observed item scores, or through the likelihood function. Most parametric person-fit statistics use the likelihood function to derive their estimate of probability for a given response set (Meijer & Sijstma, 1995). Two of the most well-known parametric person-fit statistics are the *Caution Index*, and the l_z index.

The Caution Index

The Caution Index was originally conceptualized under Student-Problem (S-P) curve theory and was applicable to either items or respondents. Tatsuoka and Linn (1983) provided an informative review of S-P curve theory: In S-P theory, a data matrix of items and respondents (with their binary response data) is created. This data matrix is arranged so that respondents are ordered from high to low total test scores in the rows, and that items were ordered from easiest to hardest in the columns. The S-curve is developed by creating vertical lines for each respondent (i.e., for each row) corresponding to the number of items they answered correctly and connecting the vertical lines. Similarly, the P-curve is created by making a mark corresponding to the number of respondents that answered an item correctly (in the columns) and connecting those marks. In this way, row and column sums and proportions can be calculated. Additionally, a "perfect" S-curve can be created by changing all values falling above the created S-curve to *1* (i.e., correct), and all values falling below the S-curve to *0* (i.e., incorrect). Given all this, the Caution Index was defined as the ratio of observed covariance between the S- and P-curves to the covariance between the S- and P-curves assuming a perfect S-curve. The equation for the caution index is given in Equation 5:

$$C_j = 1 - \frac{\sum_{i=1}^{n} \left(y_{ji} - p_j\right)\left(y_i - p\right)}{\sum_{i=1}^{n} \left(M_{ji}^{S} - p_j\right)\left(y_i - p\right)}, \quad (5)$$

where y_{ji} is the binary response of person j to item i, p_j is the proportion of correct responses for person j, y_i is the item (i.e., column) sum for item i, p is the proportion of correct responses across the entire data matrix, and M_{ji}^{S} is the binary response of person *j* to item i, assuming that the responses came from a data matrix with a perfect S-curve.

The Caution Index was not related to IRT until Tatsuoka and Linn (1983) presented five extended caution indices (ECI), which demonstrated that the S-curve could be conceptualized as a discrete test response curve. The simplest ECI conversion (referred to as ECI_1 and not presented here) merely replaces the $(M_{ji}^{S} - p_j)$ term from Equation 5 with an IRT equivalent term: $[S_{\hat{\theta}_j}(\hat{b}_i) - T(\hat{\theta}_j)]$, where, $\hat{S}_{\theta_j}(\hat{b}_i)$ is the estimated person response function for the estimated difficulty of item i, and $T(\hat{\theta}_j)$ is the test response function at the estimated theta (i.e., ability) of person j. Of the four remaining ECI conversions, ECI_2 and ECI_4 are the two that have been researched most often (Sinharay, 2016). ECI_2 is a further extension of ECI_1 given by Equation 6:

$$ECI_{2_j} = 1 - \frac{\sum_{i=1}^{n} \left(y_{ji} - p_j\right)\left[G(\hat{b}_i) - G\right]}{\sum_{i=1}^{n} \left[S_{\hat{\theta}_j}(\hat{b}_i) - T(\hat{\theta}_j)\right]\left[G(\hat{b}_i) - G\right]}, \quad (6)$$

where the new term, $[G(\hat{b}_i)\text{-}G]$, represents the group response function at the estimated difficulty for item i minus the average of the group response function (G). ECI_4 is another small extension on ECI_2. ECI_4 is given in Equation 7:

$$ECI_{4_j} = 1 - \frac{\sum_{i=1}^{n}\left(y_{ji} - p_j\right)\left[S_{\hat{\theta}_j}(\hat{b}_i) - T(\hat{\theta}_j)\right]}{\sum_{i=1}^{n}\left[G(\hat{b}_i) - G\right]\left[S_{\hat{\theta}_j}(b_i) - T(\hat{\theta}_j)\right]}, \quad (7)$$

in which $[S_{\hat{\theta}_j}(\hat{b}_i) - T(\hat{\theta}_j)]$ and $\left[G(\hat{b}_i) - G\right]$ have been swapped. Equation 7 represents a covariance ratio of the relationship between an individual's response (y_i) and the estimated person response vector at a given theta ($S_{\hat{\theta}_j}$; the numerator) divided by the covariance of the group response curve minus the test response curve and the person response curve minus the test response curve at a given level of estimated theta (the dominator).

Practically, ECI_2 provides the covariance of an individual and the overall group response curve. In this way, ECI_2 can be seen as a comparative statistic; an individual response pattern is compared to the normed group. On the other hand, ECI_4 compares an individual's response pattern to the person response curve at a given level of theta: how well the response pattern fits the model suggested curve (Tatsuoka & Linn, 1983). Tatsuoka (1984) suggested the standardization of these caution indices to address the fact that the original ECI statistics resulted in inflated values at extreme levels of theta. Additionally, Tatsuoka successfully used ECI_2 and ECI_4 to measure student misconceptions on an achievement test. However, they are generally outperformed by other person-fit statistics when they have been compared (Karabatsos, 2003; St-Onge et al., 2011; Tendeiro & Meijer, 2014). However, recent research has suggested that ECI_4 might be as effective as other person-fit statistics in certain situations (Sinharay, 2017).

The l_z Index

Drasgow et al. (1985) initially proposed the l_z index as a method of examining nonresponse. Since then, it has gone through several iterations. The popularity of the l_z index is partly because it is based on the likelihood function, which makes it relatively easy to understand compared to other parametric person-fit statistics. The general form of the likelihood function is given in Equation 8:

$$\mathrm{L}(\boldsymbol{\theta}|\mathbf{y}) = \prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{f}(y_{\mathrm{i}}|\boldsymbol{\theta}), \tag{8}$$

where $\boldsymbol{\theta}$ is a vector containing parameter values, and $\mathbf{y}$ is a vector containing observed responses. The likelihood function is easy to apply to random samples, and it provides a method for examining what parameter values result in a higher likelihood that the observed values were obtained. Additionally, the likelihood function can be easily extended to measurement (and other) contexts (Bilder & Loughin, 2015).

Drasgow et al. (1985) retrofitted an older person-fit statistic which also used the likelihood function, l_o (Levine & Rubin, 1979), to construct their new l_z index. The l_o statistic is given in Equation 9:

$$l_o = \sum_{\mathrm{i}=1}^{\mathrm{n}} y_{\mathrm{i}} \left[\log W_{\mathrm{i}}(\hat{\theta}_{\mathrm{d}})\right] + (1 - y_{\mathrm{i}})\left[\log Q_{\mathrm{i}}(\hat{\theta}_{\mathrm{d}})\right], \tag{9}$$

where y_{i} is the correct/incorrect item score, W_{i} is the item characteristic curve of the correct option for item i, $\hat{\theta}_{\mathrm{d}}$ is the maximized likelihood function of the dichotomous model and $Q_{\mathrm{i}} = 1 - P_{\mathrm{i}}(\hat{\theta}_{\mathrm{d}})$. The l_o statistic is then used to obtain the l_z index as seen in Equation 10:

$$l_z = \frac{\left[l_o - E_3\left(\hat{\theta}_d\right)\right]}{\sigma_3\left(\hat{\theta}_d\right)}, \tag{10}$$

where E_3 is the conditional mean of the three-parameter logistic (3PL) IRT model and σ_3 is the conditional standard deviation of the 3PL model. Drasgow et al. also proposed a more general version of the l_z index: z_h. Though, the only major difference between l_z and z_h is that means and standard deviations from a 3PL model with independent item responses (a histogram model, in their terminology) are used to compute the final statistic.

Since the l_z statistic is contingent on the likelihood function, which itself is dependent on fitting an IRT model (in this context), problems arise when model parameters are not known. Specifically, using an estimate of ability ($\hat{\theta}$) results in a conservative l_z result (Snijders, 2001). Snijders (2001) proposed a standardized version of l_z to partially address this issue. His standardized version of the log-likelihood-based statistic (often called l_z*) is given in Equation 11:

$$l_z^* = \frac{\sum_{i=1}^{n}\left[y_i - W_i\left(\hat{\theta}_n\right)\right]\left[\log\left(W_i\left(\hat{\theta}\right)/1 - W_i\left(\hat{\theta}\right)\right)\right]}{n^{\frac{1}{2}}\sigma_n\left(\hat{\theta}_n\right)} \tag{11}$$

where y_i is the dichotomous response to item i, W_i is the item characteristic curve of item i, and $n^{1/2}_{\sigma_n}\left(\hat{\theta}_n\right)$ corrects the reduced variance that results from using an estimate of ability (the full proofs can be found in Snijders, 2001). Magis et al. (2012) reformulated Equation 11 to be more similar to the original l_z equation (Equation 10); This reformulation is given in Equation 12:

$$l_z^* = \frac{l_0(\hat{\theta}) - E\left[l_0(\hat{\theta})\right] + c_n(\hat{\theta}) r_0(\hat{\theta})}{\widetilde{V}\left[l_0(\hat{\theta})\right]^{1/2}}, \quad (12)$$

where r_0 is the derivative of an informative or noninformative prior (which depends on which model and estimator is used), c_n is a weight function, and $\widetilde{V}[l_0(\hat{\theta})]^{1/2}$ is the approximate standard deviation of the l_0 function. Conceptually, l_z* is simply a mean and variance adjusted version of l_z that attempts to correct the conservativeness of l_z when using estimated theta (Magis et al., 2012).

While l_z is easy to understand and apply, research has shown that cautious application of the l_z statistic is necessary. Specifically, it may not perform well when tests or surveys do not have item difficulties that cover the ability range of the population being tested; it may be less powerful depending on the length of the test or survey; it is not well suited for use in computerized adaptive tests; it does not necessarily outperform the nonparametric person-fit statistics; and it requires the underlying IRT model to fit the data well (Armstrong et al., 2007; Meijer & Tendeiro, 2012; van Krimpen-Stoop & Meijer, 1999). Several of these issues carry over to l_z*, and research has shown that there is often not much difference in l_z and l_z* (Meijer & Tendeiro, 2012; van Krimpen-Stoop & Meijer, 1999).

Nonparametric Person-Fit Statistics

Recall that most person-fit statistics either compare an individual to a measurement model or to a group to assess aberrant response patterns. In general, nonparametric person-fit statistics will do the latter: they compare an individual's responses to those of a group to determine if they made responses that vary from the

norm. However, a few nonparametric person-fit statistics also assess fit based on an underlying nonparametric IRT (NIRT) model (Meijer & Sijtsma, 1995). As with all nonparametric statistics, nonparametric person-fit statistics do not assume an underlying distribution to data when applied. Having no (or few) specific requirements based on an underlying distribution allows nonparametric statistics to be applied more flexibly than their parametric counterparts. Some have argued this flexibility makes nonparametric statistics particularly suited for use in the social sciences (Siegel, 1956). Interestingly, nonparametric person-fit statistics have been shown to perform similarly when compared to their parametric counterparts and even outperform them in some situations (Dimitrov & Smith, 2006; Emons, 2008; Karabatsos, 2003; Niessen et al., 2016; St-Onge et al., 2011; Tendeiro & Meijer, 2014). While there are many nonparametric person-fit statistics, only the most well-known (and by extension, well studied) statistics will be discussed: Guttman errors, the *U3* statistic, and the H^T statistic.

Guttman Errors

Guttman errors are the simplest form of nonparametric person-fit statistic used to examine aberrant responses. They stem from the Guttman Scalogram, in which items were arranged from easiest to hardest difficulty. In a perfect Guttman scale, participants would respond correctly up to the point where an item is too difficult for them, and then would be unable to answer any further items correctly. A perfect Guttman scale would also result in distinct response patterns, for example: *11100* and *11000* would represent two response patterns of individuals who were able to answer 3 and 2 items correctly out of 5, respectively. In this way, respondents could easily be compared in terms of their ability on the construct being measured.

Guttman's scale is an interesting measurement model, but it has fallen out of use as it is considered a "pre-statistical" measurement practice (Proctor, 1970). In evaluating the appropriateness of Guttman Scalogram to data, Guttman introduced the Guttman error: the situation in which a respondent answers an easy item incorrectly and a more difficult item correctly. In this way, a Guttman error represents a deviance from what normally would be expected by a Guttman pattern. Even in the absence of any specific measurement model a respondent should not have many Guttman errors.

The simplest way to use Guttman errors to assess response patterns is to obtain a count. Assuming that *k*-items are arranged from easiest to hardest (e.g., via proportion correct or IRT *b*-parameters), the number of Guttman errors can be obtained from Equation 13,

$$G = \sum_{\mathrm{i}=1}^{k-1} \sum_{\mathrm{g}=\mathrm{i}+1}^{k} f_{\mathrm{ig}} \tag{13}$$

where f_{ig} represents a Guttman error for items i and g (1 denotes a Guttman error, while 0 denotes no Guttman error for the item pair). Meijer (1994) realized that *G* could be confounded with the number-correct score. To address this issue, Meijer proposed G^*, which is given in Equation 14:

$$G^* = \frac{G}{X_{\mathrm{j}}\left(k - X_{\mathrm{j}}\right)}, \tag{14}$$

where *G* refers to the result of Equation 13 (or the number of Guttman errors), X_{j} refers to an individual's number-correct score, and *k* refers to the number of items on the test or survey. Essentially, the denominator is the maximum possible number of Guttman errors

for a given number-correct score. Early research on the number of Guttman errors showed that it performed similarly to other nonparametric person-fit statistics in certain conditions (Harnish & Linn, 1981; Meijer, 1994). However, more recent research investigating Guttman errors for identifying aberrancy has been mixed. Some research has shown that Guttman errors are outperformed by other parametric (l_z) and nonparametric (*U3* and H^T) person-fit statistics, particularly when used to examine polytomous items (Beck et al., 2019; Karabatsos, 2003), while other research suggests that Guttman errors perform well in a variety of conditions (Emons, 2008; Niessen et al., 2016). Interestingly, the other nonparametric person-fit statistics discussed below can be conceptualized in terms of Guttman errors.

U3

Van der Flier (1982) proposed a statistic he dubbed U‴, which was designed to assess the deviance of response vectors. In this case, a deviant response vector refers to a response vector that is less probable given estimated item difficulties. While this sounds quite similar to methods employed by parametric person-fit statistics, deviance scores only allow for the ordinal (i.e., ranked) assessment of the underlying probabilities. The U''' statistic is given in Equation 15:

$$U'''(X) = \frac{\log(Pmax) - \log(\mathrm{P}(Z))}{\log(Pmax) - \log(Pmin)}, \tag{15}$$

where P(Z) is the probability of response pattern *Z*, and *Pmax* and *Pmin* represent the probabilities of the most and least deviant response patterns that could result in the same number correct score, respectively. Using this equation will result in a $U''' = 0$ if the pattern is not at all deviant (essentially, a response vector with no Guttman errors), and a

U''' = 1 if the pattern is as deviant as possible (essentially, a response vector with the maximum amount of Guttman errors). The naming convention of *U3* began as more research was conducted on U''' along with a more general form of the equation. The more general form of *U3* is given in Equation 16:

$$U3(X) = \frac{\sum_{i=1}^{Y_+} \log\left(\frac{\pi_i}{1 - \pi_i}\right) - \sum_{i=1}^{I} \mathbf{y}_i \log\left(\frac{\pi_i}{1 - \pi_i}\right)}{\sum_{i=1}^{Y_+} \log\left(\frac{\pi_i}{1 - \pi_i}\right) - \sum_{i=I-Y_++1}^{I} \log\left(\frac{\pi_i}{1 - \pi_i}\right)}, \tag{16}$$

where I is a set of dichotomous items, $\mathbf{y}_i$ is the binary response score vector, and $Y_+ = \sum_{i=1}^{I} \mathbf{y}_i$. Finally, let π_i be the proportion of correct responses to item i in the population. For a given Y_+, all terms will be constant except for $\sum_{i=1}^{I} \mathbf{y}_i \log\left(\frac{\pi_i}{1 - \pi_i}\right)$. As Equation 16 is a generalized form of Equation 15, results from the two equations will be identical. Just as with Guttman errors, *U3* was found to be confounded with the number correct score. Van der Flier (1982) standardized *U3* to address this issue. The standardized form of *U3*, *ZU3*, is given in Equation 17:

$$ZU3 = \frac{U3 - E(U3)}{[Var(U3)]^{1/2}}, \tag{17}$$

where $E(U3)$ and $Var(U3)$ are the expectation and variance of *U3*, respectively. For the full derivation of $E(U3)$ and $Var(U3)$, please see van der Flier (1982). The *U3* and *ZU3* statistics have been successfully used to identify aberrancy in certain conditions (van der Flier, 1982; Emons et al., 2005). However, more research has shown that *U3* is generally outperformed by other person-fit statistics and there is some question about the

practicability of its assumptions (Beck et al., 2019; Emons et al., 2002; Emons, 2008; Karabatsos, 2003; Meijer, 1994; St-Onge et al., 2011).

H^T

H^T is the nonparametric person-fit statistic that has shown the most promise for detecting aberrant response patterns. Coefficient H^T owes its origin to Mokken (1971), who proposed a statistic called the *coefficient of scalability* (*H*) as a method to assess the quality of a unidimensional scale under his proposed nonparametric IRT models (Mokken's work was also an extension of Loevinger's [1948] work). Mokken proposed two models: the monotone homogeneity model and double monotonicity model (Mokken, 1971; Sijstma & Molenaar, 2002). Under the monotone homogeneity model, total test scores can be treated as ordinal measures of latent ability if: 1) the test is unidimensional, 2) the responses to items are locally independent, and 3) the item response curves are nondecreasing (i.e., monotonic). The double monotonicity model makes these same three assumptions but adds a fourth: that the item characteristic curves are nonintersecting. Coefficient *H* has been applied to these models in two ways: 1) to determine how closely a test follows a perfect Guttman Scale, or more practically, 2) to determine if an item set (i.e., scale) has enough information in common to be considered unidimensional (Sijstma & Molenaar, 2002). The coefficient of scalability can be obtained at the item (H_i; often called the item scalability coefficient) or test (*H*) level.

Coefficient H^T was originally proposed as a method for evaluating the last assumption made by the double monotonicity model: the non-intersection of item characteristic curve (ICC; Sijstma & Meijer, 1992). Obtaining H^T was as simple as

transposing the standard data matrix and calculating *H*. Coefficient *H*, and by extension H^T, is given in Equation 18:

$$H = 1 - \frac{\sum_i \sum_{g \neq i} \left(P_{ig} - P_i P_g \right)}{\sum_i \sum_{g>i} P_i \left(1 - P_g \right) + \sum_i \sum_{g<i} P_g (1 - P_i)} , \tag{18}$$

where P_i, P_g, and P_{ig} represent the probability of a correct answer on item i, g, and both items i and g, respectively, and it is assumed that $P_i \leq P_g$. Since coefficient H^T is calculated using a transposed data matrix, item i and g would become person i and g. Equation 18 can also be expressed as a more explanatory covariance ratio, which is given in Equation 19:

$$H = \frac{\sum_i Cov\left(y_i, R_{(i)} \right)}{\sum_i Cov_{\max}\left(y_i, R_{(i)} \right)}, \tag{19}$$

where y_i is the response for item i, and $R_{(i)}$ is the *rest score*: the total correct score for all items except item i. $\sum_i Cov_{\max}\left(y_i, R_{(i)} \right)$ represents the maximum covariance between the response to item i and the rest score; it is obtained by correcting any Guttman errors that occur in the data used for this calculation. Recall that *H* applies to an entire test and was originally used for test construction and the assessment of quality. When applied to a transposed data matrix (H^T), the numerical result indicates to what extent the item characteristic curves intersect: low values of H^T suggest more intersection while high values suggest there is less intersection (Sijstma & Meijer, 1992).

While *H* and H^T are useful, H_i has more applications in the removal of aberrant response patterns. Recall that H_i is the item scalability coefficient. It was originally used as a method for item selection by Mokken (1971). As with *H*, H_i assesses the

homogeneity of a specific item with the set of remaining items. Coefficient H_i is given in Equation 20:

$$H_i = 1 - \frac{\sum_{g \neq i}\left(P_i - P_{ig}\right)}{\sum_{g>i} P_i\left(1 - P_g\right) + \sum_{g<i} P_g\left(1 - P_i\right)}, \tag{20}$$

where all terms are interpreted the same as in Equation 18. Just as with *H*, Equation 20 can also be expressed more clearly with a covariance ratio, which is given in Equation 21:

$$H_i = \frac{Cov\left(y_i, R_{(i)}\right)}{Cov_{\max\left(y, R_{(i)}\right)}}, \tag{21}$$

where all terms are interpreted the same as in Equation 19. Similar to how *H* is calculated on a transposed data matrix to obtain H^T, H_i can be calculated on a transposed data matrix to obtain H_i^T. Coefficient H_i^T then represents the "scalability" of a person when compared to all other people in a given data set. In this way, person i can be compared to the rest of the sample to determine if his or her response pattern is homogenous with the other response patterns (Sijtsma & Meijer, 1992; Sijtsma & Molenaar, 2002).

Research regarding H_i^T as an index of aberrant responding has been largely successful. Coefficient H_i^T is often shown to outperform other nonparametric and parametric person-fit statistics in a variety of conditions (Beck et al., 2019; Dimitrov & Smith, 2006; Karabatsos, 2003; Tendeiro & Meijer, 2014). However, some studies have shown that H_i^T does not perform as well when identifying cheating, and that H_i^T may perform similarly to some parametric person-fit statistics in certain contexts (St-Onge et

al., 2011; Sinharay, 2017). Additionally, Beck et al. (2019) found that the effectiveness of H_i^T varied widely when predicting a proxy for careless responding in polytomous data.

The Polytomous Problem

Previously, the discussion about person-fit statistics has largely focused on their application in dichotomous contexts. Unfortunately, the usage of person-fit statistics with polytomous items requires that the person-fit statistic be generalized for that application. While the polytomous generalization of a person-fit statistic might not sound very complicated, not many person-fit statistics have been generalized in this way. However, it could be argued that person-fit statistics would be *more useful* applied to polytomous contexts.

Consider the application of person-fit statistics in dichotomous contexts: a test resulting in binary responses is often going to be an achievement test. While cheating is a concern on achievement tests (and a useful application for person-fit statistics), many of the other response biases are not (e.g., SDR, acquiescence, CR, etc.). In contrast, consider a typical example of a low-stakes, polytomous survey: a researcher investigating certain attitudes in a sample of undergraduate students or in an online sample. In this situation, the researcher needs to be concerned with a bevy of response biases. If person-fit statistics are useful indicators of aberrant response patterns, they are almost mandatory for use in the polytomous example. Despite the apparent usefulness of person-fit statistics in polytomous contexts, only a handful of researchers have attempted to apply them in such a manner. The person-fit statistics that have been investigated most often in polytomous contexts are Guttman errors, *U3*, H_i^T, and l_z (Beck et al., 2019; Emons, 2008).

In dichotomous contexts, recall that a Guttman error is defined as an individual getting an easy item incorrect, but responding correctly to a harder item. To understand how Guttman errors work in a polytomous context, a foundational concept from polytomous nonparametric NIRT must be briefly discussed. In polytomous NIRT an item score is defined by Equation 22:

$$y_{ji} = \sum_{h=1}^{m-1} V_{jih}\,, \tag{22}$$

where, y_{ji} is the score of person j on item i, m represents the number of response categories (ordered as *0, 1, …, m*-1), assuming m is equal for all items, and V_{jih} is the decomposition of the m-1 dichotomous item steps, where h is ordered as 1,2, ..., m-1. If X_{ji} is greater than h, V_{jih} will equal 1, otherwise, V_{jih} will equal 0 (Mokken, 1997; Molenaar, 1997). Conceptually, X_{ji} is the number of item steps that person j passed on item i, where an *item step* is defined as having enough of the underlying latent trait to move from one response category to another (e.g., moving from response category *0* to response category *1*, assuming higher categories relate to a higher level of the underlying latent trait). In this vein, $\hat{\pi}_{ih}$ can be defined as the proportion of individuals who passed item step h for item i in a particular sample. Similar to *p*-values in CTT, $\hat{\pi}_{ih}$ can be treated as a measure of item step difficulty, with higher proportions suggesting an easier item step.

There is now a basis to define Guttman errors in a polytomous context. Conceptually, a polytomous Guttman error occurs when an individual passes a difficult item step (i.e., an item step with a small $\hat{\pi}_{ih}$), but not an easier item step. Polytomous

Guttman errors are assessed for **every** item step pair on a given survey, and the higher prevalence of polytomous Guttman errors should equate to more misfit (Emons, 2008).

Recall that Guttman errors are the simplest person-fit statistic used to assess aberrant response patterns in dichotomous data. Guttman errors are still the simplest person-fit statistic used for polytomous contexts, but the generalization has resulted in a somewhat complex method for assessing aberrancy. For example, consider a 3-item survey that uses a 5-point, Likert-type response format. For each individual completing this simple survey, 105 item step pairs would need to be assessed for Guttman errors. Regardless of the complexity added by polytomous data, the number of Guttman errors has been shown to identify aberrancy as well as, or better than, other statistics in certain conditions (Emons, 2008; Niessen et al., 2016). However, others have found that Guttman errors do not perform very well in polytomous contexts (Beck et al., 2019).

While the nonparametric person-fit statistics can all be conceptualized in terms of Guttman errors, fitting the parametric l_z index to polytomous contexts is conceptually easier, but practically more difficult. Recall that l_z is assessing the likelihood of a given response pattern given an underlying measurement model. In order for l_z to be applied to polytomous data, all that is required is to define a set of polytomous IRT model parameters. However, parametric IRT models can be quite restrictive, and appropriate model fit remains key to accurately identifying misfitting persons (Meijer & Baneke, 2004; Meijer & Tenderio, 2012). Meijer and Baneke (2004) offered several reasons that parametric IRT is often not a good fit for noncognitive measurement (i.e., measurement of attitudes, personality, frequency of behaviors, etc.). They argued that parametric IRT imposes a specific structure on the data that is often not reflective of how the survey was

constructed, and that it requires large sample sizes that are often unrealistic for noncognitive contexts. Both of these criticisms can be solved by using nonparametric IRT approaches to assess misfit. In addition to these reasons, the most commonly used parametric person-fit statistic, l_z, has been shown to perform similarly, or worse, to nonparametric person-fit statistics in polytomous contexts (Beck et al., 2019; Emons, 2008; Niessen et al., 2016).

While nonparametric person-fit statistics are often touted as equivalent to, or more useful, than parametric person-fit statistics in dichotomous data, few studies have investigated their usefulness in polytomous data. Recall that H_i^T and *U3* can also be conceptualized in terms of Guttman errors: where H_i^T is a ratio of observed covariance and the maximum covariance (i.e., the covariance if there had been no Guttman errors). Additionally, *U3* = 0 is equivalent to a response vector being a Guttman vector (i.e., a vector of responses with no Guttman errors; Meijer, 1994). Both *U3* and Coefficient H_i^T can still be conceptualized this way when they have been generalized for polytomous items. Additionally, H_i^T has been shown to test the assumption of nonintersecting ICCs made by the double monotonicity model in polytomous contexts accurately (Ligtvoet et al., 2010).

However, when Beck and colleagues (2019) applied H_i^T, *U3*, and the number of normed Guttman errors to polytomous survey data, their effectiveness at identifying aberrancy was variable. They found that Guttman errors and *U3* performed poorly overall. Additionally, they found that H_i^T performed better, but it was still below expectations based on studies where H_i^T was applied to dichotomous data. It should be noted that Beck and colleagues were using the three statistics to predict a proxy for

inattentive responding, but it was unclear why the person-fit statistics performed so variably in the polytomous context. Coefficient H_i^T improved model fit when using a cutoff that minimized false negatives and positives (similarly or better to other person-fit statistics) but performed poorly when a cutoff that offered the greatest discrimination (i.e., area under the curve from a receiver operating characteristic curve analysis) was used. Additionally, both Guttman errors and *U3* performed poorly in terms of area under the curve from a ROC analysis and on improvement to model fit. It is unclear why H_i^T, *U3*, and the number of Guttman errors showed mixed results using these cutoffs in polytomous data, as such cutoffs had been used in the past to some success (Karabatsos, 2003). The findings of Beck et al. (2019) have raised questions about potential limitations when applying nonparametric person-fit statistics in polytomous contexts.

The Current Study

The current study investigated the efficacy of H_i^T, *U3*, and the number of Guttman errors for the detection of aberrant responses in polytomous, noncognitive contexts. Specifically, these person-fit statistics were investigated under a variety of simulated conditions to determine if, when, and how they should be applied to detect aberrancy in polytomous response patterns. These conditions included type of response bias (i.e., aberrancy), number of response options, dimensionality, and test length. The type of response bias, test length, and the number of response options have been previously shown to impact the detection ability of nonparametric person-fit statistics (Emons, 2008; Karabatsos, 2003; Sinharay, 2017; St-Onge et al., 2011; Tendeiro & Meijer, 2014). However, the impact of the number of response options on H_i^T has not been investigated.

Additionally, the impact of dimensionality on these person-fit statistics in polytomous contexts has not been investigated.

Multidimensionality represents a potential problem for nonparametric person-fit statistics, as they are developed and applied under basic NIRT models (i.e., the Monotone Homogeneity Model and the Double Monotonicity Model; Sijtsma & Molenaar, 2002). The first assumption of these models is unidimensionality. Some research has shown that fitting a unidimensional IRT model to multidimensional data does not result in biased $\hat{\theta}$ or item parameter estimates, but the precision of $\hat{\theta}$, item discrimination, and item difficulty estimates decrease as the severity of violations increase (Crișan et al., 2017). Additionally, it has been shown that parametric IRT models are robust to violations of unidimensionality, particularly in the presence of a strong general factor, but can lead to an increase in root mean square errors in the item discrimination and item difficulty parameters (De Ayala, 1994; Drasgow & Parsons, 1983; Harrison, 1986). While the nonparametric person-fit statistics have been generalized for use in polytomous data, the impact of multidimensionality on nonparametric person-fit statistics, or on unidimensional NIRT models, has not been investigated. To this end, the current study will make five specific hypotheses:

- Hypothesis 1) H_i^T will show greater sensitivity, specificity, positive predictive values, and negative predictive values when compared to *U3* and the number of Guttman errors.
- Hypothesis 2) Within person-fit statistics, the sensitivity, specificity, positive predictive values, and negative predictive values will be similar across aberrancies.

- Hypothesis 3) All person-fit statistics will show increased sensitivity, specificity, positive predictive values, and negative predictive values as the number of response options increases.
- Hypothesis 4) All person-fit statistics will show decreased sensitivity, specificity, positive predictive values, and negative predictive values as the dimensionality of data increases.
- Hypothesis 5) All person-fit statistics will show increased sensitivity, specificity, positive predictive values, and negative predictive values as test length increases.

CHAPTER III: METHOD

This study utilized four simulation factors: 1) type of response bias, 2) number of response options, 3) test length, and 4) dimensionality. For the type of response bias condition, six response biases were modeled: disacquiescence, acquiescence, midpoint responding, extreme responding, social desirability responding, and careless responding. For the number of response options condition, four response options of differing lengths were simulated: four, five, six, and seven. These four response option conditions were based on empirical research regarding the optimum and most common number of response categories for noncognitive measures (DeCastellarnau, 2018; Revilla et al., 2014; Weijters et al., 2010). Test length was simulated as short (using 12 items) and medium (using 36 items). Finally, four different conditions of dimensionality were tested: unidimensional, two-dimensional, three-dimensional, and four-dimensional. These dimensions will be correlated factors from a single scale. The result is a simulation study with 4 x 4 x 2 x 6 conditions, for a total of 192 unique combinations. R (R Core Team, 2021) was used for all item generation procedures and analyses.

Simulation Procedures

Following procedures in the extant literature for generating realistic polytomous item response data that accounts for dimensionality, the Multidimensional Graded Response Model (MGRM) was used (Bulut & Sünbül, 2017; De Ayala, 1994; Jiang et al., 2016). One expression of the MGRM is given in Equation 23:

$$P_{jk}(\boldsymbol{\theta}) = \frac{1}{1 + \exp\left[-D\sum_{h=1}^{H}\left[a_{jh}\left(\theta_h - b_{jk}\right)\right]\right]}, \tag{23}$$

where $P_{ik}(\boldsymbol{\theta})$ is the probability that a respondent with a latent trait vector of length H ($\boldsymbol{\theta}$) will respond in category k (with ordered categories of k + 1) or higher for item i; D is an optional scaling constant of 1.702 (otherwise $D = 1$); a_{ih} is the item discrimination parameter of item i on dimension h; θ_h is the latent trait of interest for dimension h; and b_{ik} is the item boundary parameter for category k on item i. By default, the probability of responding in the lowest category or higher is defined as 1 ($P_{i0}(\boldsymbol{\theta}) = 1.0$), and the probability of responding in the highest category is defined as 0 ($P_{i(k+1)} = 0.0$).

To simulate the measurement model, an instrument with four intercorrelated factors was used as a basis for the simulation conditions. (See Beck, 2015 for an overview, and the general factor structure for the simulation is provided in Figure B1). A multivariate normal distribution was used to simulate 1000 latent trait vectors across four dimensions. Each of the four dimensions was assigned a mean ranging from -.001 to .092, and a covariance matrix with off-diagonal terms ranging from 0.31 to 1.11 was used to define the covariance structure of these data. The means and the covariance matrices were adapted from parameters obtained from Beck (2015) and are provided in Table A1.

Item Characteristics

Item boundary parameters (b_{ik}) were based on values obtained from fitting data from Beck (2015) to the Graded Response Model using the *ltm* package (Rizopoulos, 2006) in R. Recall that the number of response options was a manipulated condition for this study, and that the number of response options generated were four, five, six, and seven. For each of these four conditions, the item boundary parameters were created from appropriate quantiles (i.e., terciles for $k = 4$, quartiles for $k = 5$, and quintiles for $k = 6$) using the item boundary parameter estimates obtained from Beck (2015). These quantiles

formed the basis of the uniform distributions used to obtain the item boundary parameters for each condition. The uniform distributions used to create the item boundary parameters are given in Table A2.

For the 144 total items across four dimensions (36 loading primarily on each dimension), item discrimination parameters were generated from unique uniform distributions. The range for these uniform distributions were taken from the ranges of item discrimination parameters obtained by fitting data from Beck (2015) to the Graded Response Model using the *ltm* package (Rizopoulos, 2006) in R. The uniform distributions that the primary item discrimination parameters were drawn from were: $a1_{\text{primary}} \sim U(1.895, 3.296)$; $a2_{\text{primary}} \sim U(2.657, 4.668)$; $a3_{\text{primary}} \sim U(2.851, 6.651)$; and $a4_{\text{primary}} \sim U(1.803, 3.837)$. To ensure the data were realistic as possible, each set of 36 items had nonzero item discrimination parameters for all dimensions. Similar to procedures used by Finch (2011), smaller item discrimination parameters were generated for the nonprimary dimension loadings. These smaller item discrimination parameters were randomly drawn from a uniform distribution ($a_{\text{secondary}} \sim U[0.1, 0.4]$) for all dimensions. The full table of item discrimination parameters used to generate data is provided in Table A3. After the latent trait vector (discussed in the previous section), item discrimination parameters, and item boundary parameters were created, simulated responses to all items were generated under the MGRM model using the *mirt* (Chalmers, 2012) package in R.

Data Set Generation

For each unique dimensionality, number of response options, and test length condition (4 x 4 x 2), a data set was created by randomly sampling items from the

appropriate dimension(s). The sampling occurred such that there was an equal number of simulated responses to items from each dimension represented in multidimensional data sets, and these responses were sampled without replacement. This sampling procedure resulted in 32 static data sets that were used for all further analyses. These static data sets each contained 12 or 36 items and had items loading on between one and four dimensions with an equal number of items coming from each dimension. The full specifications of these data sets are provided in Table A4. After these 32 data sets were generated, any individual response vectors with zero variance were removed as they interfered with the calculation of some nonparametric person-fit statistics. Additional simulated observations were removed at random until all data sets had 900 simulated respondents. To ensure that these static data sets were comparably reliable, coefficient alpha and Omega Total (McDonald, 1999) were investigated for each data set within the number of response option conditions. Across dimensions, all reliability estimates fell within a (±) .05 range.

Adding Aberrancy

For each data set, 100 aberrant response vectors were added for a total of 1000 simulated respondents per data set. Adding aberrancy in this manner created data sets where exactly 10% of all simulated respondents engaged in aberrant responding. Recall that the extant literature suggests that ten percent of respondents are expected to engage in aberrant responding in a given noncognitive survey sample (Beck et al., 2019; Johnson, 2005; Maniaci & Rogge, 2014; Mckibben & Silvia, 2015; Meade & Craig, 2012). The process of adding aberrancy was repeated 1000 times for each condition, with aberrant respondents being generated from a new multivariate normal distribution every

iteration. This step resulted in the creation of 32,000 data sets (4 x 4 x 2 conditions x 1000 replications) for each aberrant response pattern.

Acquiescence and Disacquiescence

To simulate both acquiescence and disacquiescence, one hundred additional response vectors were generated using modified item boundary parameters. An approach similar to Emons (2008) and Rossi et al. (2001) was used to modify the item boundary parameters. Namely, the item boundary parameters were linearly transformed. Then, item responses using theta values identical to those described in the generation of the base data sets were used to generate additional aberrant response vectors. The linear transformation of the item boundary parameters for both acquiescence and disacquiescence followed a similar pattern, albeit in different directions. The linear transformation for acquiescence is given in Equation 24:

$$\delta^{*}_{ki} = \delta_{ki} + s_{\delta_i}, \tag{24}$$

where δ^{*}_{ki} is the modified item boundary parameter *k* for item i, δ_{ki} is the unmodified item boundary parameter *k* for item i, and s_{δ_i} is the standard deviation of the item boundary parameters for item i. This linear transformation increased the likelihood that the simulated respondents endorsed higher categories. The linear transformation for disacquiescence is given in Equation 25:

$$\delta^{*}_{ki} = \delta_{ki} - s_{\delta_i}, \tag{25}$$

where all terms are defined the same as they are in Equation 24. This transformation decreased the likelihood that simulated respondents endorsed higher categories.

Extreme and Midpoint Responding

To simulate both extreme and midpoint responding, additional response vectors were generated with modified item boundary parameters. The aberrancy was generated using procedures described by Emons (2008) and Rossi et al. (2001), who used the same method and equation to simulate both extreme and midpoint responding. The item boundary parameters were linearly transformed and used to generate aberrant response vectors. The linear transformation is given in Equation 26:

$$\delta_{ki}^{*} = \exp(\varepsilon) \times \left(\delta_{ki} - \bar{\delta}_{i}\right) + \bar{\delta}_{i}, \tag{26}$$

where δ_{ki}^{*} is the modified boundary parameter *k* for item i, ε is a constant representing the size of the transformation (i.e., the size of the aberrant response effect), δ_{ki} is the *k* item boundary for item i, and $\bar{\delta}_{i}$ is the mean item boundary parameter for item i. When $\varepsilon > 0$, response vectors were modified such that midpoint responses have higher endorsement probabilities. When $\varepsilon < 0$, response vectors were modified such that extreme responses have higher endorsement probabilities.

Social Desirability Responding

Simulating social desirability responding (SDR) also used a linear transformation of item boundary parameters. In fact, it used the same linear transformation seen in Equation 24. However, the linear transformation was only performed on simulated respondents with low values of theta. This method of generating SDR has a few shortcomings. Namely, it only simulates “faking good”, that is, respondents with low levels of a desired latent trait responding in a manner that suggests higher levels of the desired latent trait. Additionally, the linear transformation was identical to the linear transformation for acquiescence responding. In essence, this method assumed that the

only difference between acquiescence responding is who engages in it (i.e., individuals with certain levels of theta for SDR), and the motivation causing it.

To simulate SDR, an additional multivariate normal distribution was used to generate response vectors. However, this new distribution had latent means falling one standard deviation below the latent means used to create the base data sets (the latent means will be: ξ_1 = -.982; ξ_2 = -.964; ξ_3 = -.903; and ξ_4 = -.868). All other aspects of this multivariate normal distribution were kept the same as the distribution used to generate the base data sets. The new multivariate normal data set was used to generate item responses to items with item boundary parameters modified by Equation 24.

Careless Responding

Careless responding was simulated by generating 100 additional response vectors from the multivariate normal distribution used to generate the base data sets. For each new simulated respondent, a random response option replaced the respondent's raw response on the selected items. Each response option in a given response set had the same likelihood to be chosen as any other response option.

This method of simulation made the explicit assumption that careless responding is equivalent to random or uniform responding (recall that both random and uniform responding are collectively included under the careless responding term). Unfortunately, the mechanisms of careless responding are not understood well enough for a more purposeful simulation. Additionally, this method of simulating careless responding made it unlikely that response vectors with no variance were created.

Estimating and Applying Nonparametric Person-Fit Statistics

Recall that the process of adding aberrancy to each data set was repeated 1000 times, selecting random items and respondents each time. The estimation of H_i^T across all conditions was conducted using the *mokken* package in R (van der Ark, 2007, 2012). The estimation of *U3* and the number of normed Guttman errors was performed using the *PerFit* package in R (Tendeiro et al., 2016).

Coefficient H_i^T does not have an established theoretical null distribution, so specific cutoff values or statistical tests are not readily available (Mousavi et al., 2019). Similarly, there is no standardized method for determining cutoff values for *U3* or Guttman errors. Therefore, to determine whether or not the nonparametric person fit statistics were identifying response vectors as aberrant, an empirical cutoff was applied. For each statistic, this empirical cutoff was derived from a bootstrapped sample of the estimated values. Specifically, the bootstrapping procedure followed these steps:

1. Coefficient H_i^T, *U3*, and normed Guttman errors were estimated for all conditions and repetitions.
2. The distributions of the nonparametric person fit statistics were treated as populations, and randomly sampled from 1000 times with replacement. This generated 1000 samples from each of the 1000 repetitions.
3. For each of the bootstrapped samples, the value falling at the 5^{th} percentile rank was found.
4. The median of the 5^{th} percentile rank values was found and treated as the cutoff value.

Performing the bootstrapping procedure as outlined above follows recommendations by Mousavi et al. (2019). Additionally, it allowed for the nominal alpha level (Type I error rate) to be close to 0.05 (Mousavi et al., 2019). Finally, using the same cutoff criteria for each person-fit statistic allowed for comparisons of outcomes across statistics.

Estimating Efficacy

The efficacy of the nonparametric person fit statistics was determined by calculating sensitivity, specificity, negative predictive value, and positive predictive value. The results of any method, procedure, or test designed to classify individuals into a category can be described in terms of true or false positives and negatives. Consider the possible results of an individual respondent being examined for aberrant response patterns as presented in Table A5. A simple equation for sensitivity is given in Equation 27:

$$\text{Sensitivity} = \frac{\text{True Positives}}{(\text{True Positive} + \text{False Negatives})}, \tag{27}$$

where sensitivity is calculated as the proportion of individuals correctly classified as aberrant to the total number of individuals who were simulated to respond aberrantly. A simple equation for specificity is given in Equation 28:

$$\text{Specificity} = \frac{\text{True Negatives}}{(\text{True Negatives} + \text{False Positives})}, \tag{28}$$

where specificity is defined as the proportion of individuals correctly classified as not aberrant to the total number of individuals not simulated to respond aberrantly. A simple equation for positive predictive value (PPV) is given in Equation 29:

$$\text{PPV} = \frac{\text{True Positives}}{(\text{True Positives} + \text{False Positives})}, \tag{29}$$

where the positive predictive value is the ratio of true positives to all positives. PPV values range from zero to one, with one representing perfect accuracy. PPV represents the proportion of individuals correctly classified as aberrant out of all individuals classified as aberrant. Conversely, negative predictive value (NPV) is the proportion of individuals correctly classified as not aberrant out of all individuals classified as not aberrant. A simple equation for NPV is give in Equation 30:

$$\text{NPV} = \frac{\text{True Negative}}{(\text{True Negatives} + \text{False Negatives})}. \tag{30}$$

Sensitivity, specificity, positive predictive value, and negative predictive value were calculated for each of the 1000 repetitions for each condition and then averaged to provide an aggregated result for each condition. The results provided information on how well the nonparametric person fit statistics correctly detected aberrant response patterns (sensitivity), correctly rejected respondents without aberrant response patterns (specificity), and how precise the nonparametric person fit statistics were at identifying aberrancy (PPV and NPV).

CHAPTER IV: RESULTS

Recall that a 6 x 4 x 4 x 2 simulation study was conducted to examine the specificity, sensitivity, positive predictive value (PPV), and negative predictive value (NPV) of *U3*, H_i^T, and the number of normed Guttman errors when applied to polytomous survey data with aberrant responses. Specifically, six types of aberrant responses were simulated: acquiescence, disacquiescence, extreme responding, midpoint responding, socially desirable responding, and careless responding. For each of these aberrancies, response data were simulated such that each had data ranging from one to four dimensions, short (12) and medium (36) test lengths, and one of four different response categories (4, 5, 6, and 7). For each unique condition of this simulation, the specificity, sensitivity, PPV, and NPV were determined based on the classification accuracy of the person-fit statistics.

For each aberrant response behavior and person-fit statistic, Analyses of variance (ANOVAs) were calculated using the obtained specificity, sensitivity, PPV, and NPV values as the outcome variables. Dimensionality, number of response options, and test length were added to the models as main effects; all combinations of these variables were also included as interaction terms (with the highest term being a three-way interaction between dimensionality, length, and the number of response options). Weighted least squares general linear models were used to generate model parameters, which were then converted into ANOVA tables. Weighted least squares estimation via the Iteratively Reweighted Least Squares (IRLS) algorithm was used to address issues with heteroscedasticity (Cohen et al., 2003; Green, 1984). Since the sample sizes were large and the resulting statistical tests were overpowered, partial omega squared values (Keren

& Lewis, 1979) were calculated for effect size and were used to make determinations about which factors to interpret. The rules of thumb presented by Cohen (1992a) were adapted to determine the magnitude of the effects (i.e., $\omega^2_p < .13$ = small, $.13 <= \omega^2_p < .26$ = medium, and $\omega^2_p > .26$ = large). The results of the ANOVAs and the effect sizes are presented near the beginning of every section for the following chapter. After the ANOVAs are presented, the provided figures are used to interpret the results. Note that the ANOVAs and effect sizes are discussed generally in the chapter for brevity, but the complete ANOVA tables with effect size estimates are provided in Appendix C.

As the results section is quite dense, the next two paragraphs are intended to aid in orienting readers. The results section is parsed into subsections for each aberrant response pattern, which are further subdivided into each person-fit statistic. For each aberrant response pattern and person-fit statistic, the ANOVA results are discussed in a general manner, then, the highest-order term that showed a large or medium effect size is interpreted. Finally, aggregated results are presented, though, these should be interpreted with caution when in the presence of practically meaningful interactions.

In terms of the ANOVA results, the sample sizes were quite large given simulation procedure (N = 32,000 for each person-fit statistic). With such large sample sizes, the p-values returned from the ANOVA were miniscule (this issue was also encountered when attempting to review the post-hoc tests). All results for all outcomes from the study were statistically significant even using strict p-value criteria ($p < .0001$). For this reason, the effect sizes were largely used to determine whether an effect would be interpreted as meaningful. Finally, while most outcomes for the person-fit statics showed at least medium-sized main effect within each response pattern, only the highest-

order terms were interpreted. Main effects become notoriously difficult to interpret when their levels depend on levels of other factors, particularly with disordinal interactions.

While only the highest-order meaningful effects were interpreted in the results section, it is important to attend to the partial omega squared values provided in Appendix C as well as the figures provided in Appendix B (Figures B2 – B73). No effects classified as small will be discussed in the results, but the discussions of the three-way interactions across aberrancies and person-fit statistics often do not explicitly emphasize that not all meaningful effects had the same magnitude. For example, in the results for Guttman errors identifying acquiescence responding, the three-way interaction explained between 24% and 30% of the variance in the outcomes, depending on the outcome, after accounting for variance explained by the other factors. Conversely, the three-way interaction for H_{i}^{T} in acquiescence responding only explained between 4% and 18% of the variance in the outcomes, depending on the outcome, after accounting for variance explained by other predictors in the model. This suggests that the three-way interaction for Guttman errors found within acquiescence responding is more meaningful/stronger than the three-way interaction found for H_{i}^{T} in terms of variance explained. However, even in this example, it is important to remember that the three-way interactions for both Guttman errors and H_{i}^{T} were classified as meaningful based on the effect size criteria adopted by this study. Therefore, while the partial omega squared results should be considered, the results will still only interpret the highest-order terms that were found to be meaningful.

Acquiescence

Guttman Errors

The average negative predictive value (NPV), positive predictive value (PPV), sensitivity, and specificity for the number of normed Guttman errors across the simulation conditions are given in Figures B2, B3, B4, and B5, respectively. Additionally, Table A6 contains the NPV, PPV, sensitivity, and specificity aggregated across simulation conditions (dimensions, number of response categories, and test length).

Given the large sample sizes across all outcome variables (NPV, PPV, sensitivity, and specificity), all terms in the ANOVA were statistically significant. The effect sizes were more illuminating. For all outcomes, there were large or medium effect size estimates for the dimensionality, number of response options, and test length conditions. Additionally, the interaction between dimensionality and the number of response options and the three-way interaction showed at least medium effect size estimates.

While there were lower-order terms (i.e., main effects and two-way interactions) that also showed large or medium effect sizes, the results focus on the three-way interaction. Specifically, there was a large improvement in the NPV, PPV, specificity, and sensitivity of normed Guttman Errors in the 3- and 4-dimensionality conditions for seven response options on medium length tests. This effect was not seen in the other dimensionality conditions, with most other conditions performed best at 2-dimensions in the medium-length condition. Additionally, this effect was not seen in the short test condition. Four and six response options tended to perform best in the 3- and 4-dimension conditions for short tests, and the 2- and 3-dimension conditions for medium tests.

Coefficient H_i^T

The negative predictive value (NPV), positive predictive value (PPV), sensitivity, and specificity for coefficient H_i^T across the simulation conditions are given in Figures B6, B7, B8, and B9, respectively. Additionally, Table A7 contains the NPV, PPV, sensitivity, and specificity aggregated across simulation conditions (dimensions, number of response categories, and test length).

Given the large sample sizes, across all outcome variables (NPV, PPV, sensitivity, and specificity) all terms in the ANOVA were statistically significant. The effect sizes were more illuminating. For sensitivity, the number of response options and the interaction between dimensionality and test length had large effect sizes. The interaction between dimensionality and the number of response categories, the interaction between the number of response categories and length, and the three-way interaction showed medium effect sizes as well. For specificity, only the number of response options showed a large effect size; all other effect sizes were small. The effect sizes for PPV and NPV were identical to sensitivity. In Figure B9 and Table A7, seven response options consistently result in higher specificity when compared to four, six, and seven response options regardless of dimensionality or length.

In terms of sensitivity, PPV and NPV, there was a moderately sized three-way interaction. Figures B6, B7, and B8 show that while seven response options still outperform the other response options, the difference is impacted by both dimensionality and test length. All response option conditions perform their worst in the 2-dimension condition on short tests but perform their best in the 2-dimension condition on medium tests. Additionally, with six response options, there is a large increase in outcomes between short and medium tests in the 1- and 2-dimension conditions. However, the

increase is smaller in the 3-dimension condition and become a decrease in the outcomes in the 4-dimension condition.

Overall, H_i^T showed the highest outcomes at the medium test length condition when compared to the short test length, though the differences are small. In terms of the number of response categories, H_i^T performed much higher with seven response categories compared to four, five, or six response categories, particularly in terms of PPV and sensitivity. In regard to dimensionality, H_i^T performed best at 1- and 4-dimensions, with small differences existing between the two.

Coefficient *U3*

The negative predictive value (NPV), positive predictive value (PPV), sensitivity, and specificity for coefficient *U3* across the simulation conditions are given in Figures B10, B11, B12, and B13, respectively. Additionally, Table A8 contains the NPV, PPV, sensitivity, and specificity aggregated across simulation conditions (dimensions, number of response categories, and test length).

Given the large sample sizes, across all outcome variables (NPV, PPV, sensitivity, and specificity) all terms in the ANOVA were statistically significant. The effect sizes were more useful. For sensitivity, only the number of response categories showed a medium effect. There were no large or medium effect sizes for specificity. For NPV, there was a large effect for the number of response categories and a medium effect for test length. For PPV, there was a large effect for the number of response categories, and medium effects for test length, and a medium effect for the interaction between dimensionality and the number of response options.

In terms of sensitivity, Figure B12 shows five response options outperforming the other response option categories fairly consistently, with only minor differences between the other response categories across other conditions. In terms of NPV, Figure B10 shows that five response options outperformed most other response categories in most conditions, and that increasing test length mostly led to an increase in NPV. Finally, in terms of PPV, Figure B11 shows that medium-length tests offer higher PPV in most conditions compared to short tests. For the two-way interaction, for five, six, and seven response categories, there was generally an increase in PPV as dimensionality increased up to 3-dimensions, after which there was a decrease in PPV as dimensions increased from three to four. However, for five response options (the best performing response option condition), there was an increase from 1- to 2-dimensions, after which PPV started to decrease.

Overall, *U3* performed better at the medium test length condition when compared to the short test length condition, particularly when examining PPV and sensitivity. In terms of response categories, five response options noticeably outperformed four, six, and seven in terms of PPV and sensitivity. Finally, the 2- and 3-dimension conditions outperformed the 1- and 4-dimension conditions in terms of PPV and sensitivity.

Overall

For acquiescence responding, H_i^T performed the best overall. Across all conditions on average, it showed .90 Negative Predictive Value (NPV), .16 Positive Predictive Value (PPV), 8.16% sensitivity, and 95.27% specificity. While the differences between the normed number of Guttman errors and *U3* were small, Guttman errors slightly outperformed *U3*. Guttman errors showed .90 NPV, .11 PPV, 5.49% sensitivity, and

94.86% specificity aggregated across all conditions. Across all conditions, *U3* showed .90 NPV, .09 PPV, 5.12% sensitivity, and 94.80% specificity.

Disacquiescence

Guttman Errors

The negative predictive value (NPV), positive predictive value (PPV), sensitivity, and specificity for the number of normed Guttman errors across the simulation conditions are given in Figures B14, B15, B16, and B17, respectively. Additionally, Table A9 contains the NPV, PPV, sensitivity, and specificity aggregated across simulation conditions (dimensions, number of response categories, and test length).

While all ANOVA terms were significant in the hypothesis test, the effect sizes provided more information. For NPV, there were large or medium effects for: the dimensionality factor, the number of response options factor, all two-way interactions, and the three-way interaction. For both specificity and PPV, there were large or medium effects for all terms except test length. For sensitivity, there were large effects for dimensionality, the number of response options, and the interaction between dimensionality and response options.

For NPV, Figure B14 shows the three-way interaction. Most noticeably, there was little impact from any of the conditions on the five-response option condition. However, there were increases for the four-response option and seven-response option conditions at 4-dimensions on medium tests compared to short tests. Additionally, there was an increase in the NPV of six response options on medium tests compared to short tests in the 1-dimension condition. However, NPV actually decreases in the 2- and 3-dimension conditions on medium tests when compared to short tests. For specificity and PPV, there

was an increase in both outcomes for four and five response options on medium tests, except in the 3-dimension condition, which stayed roughly equivalent between short and medium tests. For six response options, there was an increase in outcomes at 1-dimension between short and medium tests. However, the outcomes of six response options started to decrease at 2-, 3-, and 4-dimensions when compared between test lengths. For sensitivity, four, five, and six response options tended to decrease from 1- to 3-dimensions and then increased at the 4-dimension condition. However, the seven-response option condition tended to increase as dimensionality increases.

Overall, Guttman errors performed slightly better in the medium test length condition when compared to the short test length condition in terms of PPV and sensitivity. In terms of the number of response categories, four response options offered the best performance with five response options performing only slightly worse. The 1-dimension condition was the best performing in terms of dimensionality, with 2-dimensions performing second best and 4-dimensions performing slightly worse than 2-dimensions.

Coefficient H_{i}^{T}

The negative predictive value (NPV), positive predictive value (PPV), sensitivity, and specificity for H_{i}^{T} across the simulation conditions are given in Figures B18, B19, B20, and B21, respectively. Additionally, Table A10 contains the NPV, PPV, sensitivity, and specificity aggregated across simulation conditions (dimensions, number of response categories, and test length).

Again, while all ANOVA terms were statistically significant, the effect sizes were used to determine medium and large effects. Sensitivity showed large or medium effects

for the number of response options, the two-way interactions between test length and dimensionality and the number of response options, and the three-way interaction. Specificity only showed large or medium effects for the number of response options. Both PPV and NPV showed large or medium effects from the number of response categories, the test length, the interaction between dimensionality and test length and between response options and test length, and the three-way interaction.

For both PPV and NPV, Figures B18 and B19 show similar results. The three-way interaction is most apparent in the four and five response option conditions. In short tests, five response options consistently performed better than four response options. However, on medium tests, four response options tended to outperform five. This difference is due to a large increase in the PPV and NPV of four response options across test lengths. However, the increase in outcomes is dependent on dimensionality. Specifically, the largest increase occurred at 1-dimension and the smallest increase in outcomes occurred at 4-dimensions. Interestingly, the 4-dimension condition is the condition in which five response options showed its largest increase between short and medium tests. Sensitivity showed a similar three-way interaction effect to NPV and PPV. Additionally, the main effect of response options can be seen on Table A10 for specificity, with four response categories outperforming all others with no deviation.

Overall, H_{i}^{T} performed better in the medium-length test condition when compared to the short test-length condition. In terms of response options, H_{i}^{T} showed the highest PPV and sensitivity with four response options. As the number of response categories increased, the PPV and sensitivity of H_{i}^{T} decreased. In terms of dimensionality, 1- and 4-

dimensions showed the highest classification accuracy for H_i^T. While 1-dimension performed the best, there were only slight differences between 1- and 4-dimensions.

Coefficient *U3*

The negative predictive value (NPV), positive predictive value (PPV), sensitivity, and specificity for *U3* across the simulation conditions are given in Figures B22, B23, B24, and B25 respectively. Additionally, Table A11 contains the NPV, PPV, sensitivity, and specificity aggregated across simulation conditions (dimensions, number of response categories, and test length).

While the ANOVA terms were all statistically significant, the effects sizes provided more information. For specificity, NPV, and PPV, all terms had medium or large effects except for the test length (i.e., the main effect of test length). For sensitivity, all terms had medium or large effects except for dimensionality and test length (again, the main effects).

For specificity, NPV, and PPV the five-response option condition noticeably outperformed the other response option conditions at 4-dimensions. However, on medium tests, both the four- and seven-response option conditions noticeably outperformed the other two at 4-dimensions. Additionally, six response options showed improvement on short tests at 2-dimensions. However, this improvement was not seen on medium length tests. For sensitivity, the effects discussed for specificity, NPV, and PPV (at least in terms of the three-way interaction) are identical.

Overall, *U3* actually performed slightly better in terms of PPV, Sensitivity, and NPV in the short test-length condition when compared to the medium test-length condition. In terms of response categories, four response categories outperformed the

others, with PPV and sensitivity decreasing as the number of response categories increased. Regarding dimensionality, *U3* performed the best in the 4-dimension condition, followed by the 2-dimension condition.

Overall

For disacquiescence responding, the normed number of Guttman errors outperformed *U3* and H_i^T when averaged across all conditions. Guttman errors showed .91 negative predictive value (NPV), .24 positive predictive value (PPV), 12.45% sensitivity, and 95.70% specificity. Coefficient *U3* performed second best, with .91 NPV, .22 PPV, 11.20% sensitivity, and 95.62% specificity. Coefficient H_i^T performed the worst, with .90 NPV, .08 PPV, 4.17% sensitivity, and 94.82% specificity.

Midpoint Responding

Guttman Errors

The negative predictive value (NPV), positive predictive value (PPV), sensitivity, and specificity for the normed number of Guttman errors across the simulation conditions are given in Figures B26, B27, B28, and B29, respectively. Additionally, Table A12 contains the NPV, PPV, sensitivity, and specificity aggregated across simulation conditions (dimensions, number of response categories, and test length).

While all ANOVA terms were significant, the effect sizes told a different story. Specifically, the only specificity showed any medium sized effects. The medium sized effects were the interaction between dimensions and response categories and the three-way interaction. Between short and medium tests, there was a decrease in sensitivity at 1- and 4-dimensions and an increase at 2- and 3-dimensions for seven response options. Additionally, there was an increase in sensitivity at 2-dimensions for five response

options between test lengths. Finally, there was a decrease between test lengths in sensitivity at 3-dimensions for four response options.

Overall, Guttman errors performed poorly when classifying individuals engaging in midpoint responding. While NPV and specificity remained stable across conditions, PPV also remained stable at values of .01 and below. In terms of sensitivity, Guttman errors showed the highest sensitivity in the 2- and 3-dimension conditions, the seven response option conditions, and on short tests. However, it should be noted that these values were never greater than 1%.

Coefficient H_i^T

The negative predictive value (NPV), positive predictive value (PPV), sensitivity, and specificity for H_i^T across the simulation conditions are given in Figures B30, B31, B32, and B33, respectively. Additionally, Table A13 contains the NPV, PPV, sensitivity, and specificity aggregated across simulation conditions (dimensions, number of response categories, and test length). Since all ANOVA terms were significant, the effect sizes were used to determine which effects were meaningful. For sensitivity, NPV, and PPV all terms showed large or medium effect sizes. However, specificity showed large or medium effect sizes for all terms except the two-way interaction between response options and test length.

In terms of the three-way interaction, there was a large increase for all outcomes between short and medium length tests for four, five, and six response categories at 1-dimension. However, this improvement diminished as the dimensionality increased. Interestingly, the 1-dimension condition actually showed a decrease in performance for seven response categories on medium tests compared when compared to short tests.

Overall, H_i^T performed better in the medium test-length condition when compared to the short test-length condition. In terms of the number of response categories, H_i^T showed the highest PPV, sensitivity, and specificity in the six-response option condition, with the seven-response option condition showing the next best classification accuracy. In terms of dimensionality, H_i^T performed noticeably better in the 1-dimension condition in terms of PPV, NPV, sensitivity, and specificity compared to the other dimensionality conditions.

Coefficient *U3*

The negative predictive value (NPV), positive predictive value (PPV), sensitivity, and specificity for *U3* across the simulation conditions are given in Figures B34, B35, B36, and B37, respectively. Additionally, Table A14 contains the NPV, PPV, sensitivity, and specificity aggregated across simulation conditions (dimensions, number of response categories, and test length).

Despite all ANOVA terms showing statistical significance, the effect size estimates provided more meaningful information. The only medium-sized effect was the interaction between dimensionality and response options for PPV. All other effect size estimates for *U3* were classified as small.

For the two-way interaction between dimensionality and response options, four and six response categories are not noticeably impacted by dimensionality. However, four and six response categories show a noticeable impact from dimensionality. Namely, four response options showed increased PPV in the 1-dimension condition compared to all other dimensions. Seven response options showed the lowest PPV at 1-dimension, which then increases at 2- and 3-dimensions, before decreasing again at 4-dimesions.

Much like Guttman errors, *U3* struggled to identify midpoint responders overall. Specificity and NPV are relatively stable across conditions at approximately 94% and .90, respectively. The PPV was also relatively constant across conditions at values of .01 or less. Finally, while there were some differences in terms of sensitivity, it should be noted that none of the values were greater than 1%. However, *U3* showed the greatest sensitivity in the 1-dimension condition with 7-response options. In terms of test length, there was only a small difference between short and medium tests (PPV = .22 and .21, respectively).

Overall

For midpoint responding, coefficient H_i^T noticeably outperformed both *U3* and the normed number of Guttman errors. Across all conditions, on average, H_i^T showed .91 negative predictive value (NPV), .26 positive predictive value (PPV), 13.29% sensitivity, and 95.83% specificity. The number of normed Guttman errors showed .90 NPV, less than .01 PPV, 0.12% sensitivity, and 94.39% specificity. Coefficient *U3* showed .90 NPV, less than .01 PPV, .22% sensitivity, and 94.41% specificity.

Extreme Responding

Guttman Errors

The negative predictive value (NPV), positive predictive value (PPV), sensitivity, and specificity for the normed number of Guttman errors across the simulation conditions are given in Figures B38, B39, B40, and B41, respectively. Additionally, Table A15 contains the NPV, PPV, sensitivity, and specificity aggregated across simulation conditions (dimensions, number of response categories, and test length).

While all ANOVA terms for all outcomes were significant, the effect sizes provided more useful information. For sensitivity and NPV, only the main effects of the number of response options and test length were medium or larger. For specificity and PPV, all terms showed a medium or larger effect size.

Table A15 provides interpretations of the main effect of response option and test length for sensitivity and NPV. Specifically, sensitivity and NPV were highest with five response categories, followed by four response categories, then six, and finally seven. For test length, medium length tests provided better sensitivity and NPV than short length tests.

For specificity and PPV the three-way interaction showed a large effect size. Figures B39 and B41 show this three-way interaction clearly. Most noticeably, there was a clear impact on the accuracy at 1-dimension for four and six response options between short and medium tests (with medium test length showing a benefit to these response option conditions). Additionally, there was a large increase in the PPV and sensitivity of seven response options on medium tests at 3- and 4-dimensions when compared to short tests. This benefit was much smaller at the 1- and 2-dimension conditions for seven response options.

Coefficient H_i^T

The negative predictive value (NPV), positive predictive value (PPV), sensitivity, and specificity for H_i^T across the simulation conditions are given in Figures B42, B43, B44, and B45, respectively. Additionally, Table A16 contains the NPV, PPV, sensitivity, and specificity aggregated across simulation conditions (dimensions, number of response categories, and test length).

With such large sample sizes, all terms in the ANOVA model were significant. However, the effect sizes showed that for sensitivity, NPV, and PPV, only the main effects for the number of response options and the test length had medium or large effects. In terms of test length, medium length tests provided better outcomes than short length tests. In terms of response categories, five response options offered the best outcomes. Four response options offered the second-best outcomes, closely followed by six response options. Seven response options showed the worst values for the outcome variables. Specificity showed no large of medium effects for any ANOVA terms suggesting that no conditions had a meaningful impact on the specificity of H_{i}^{T}.

Coefficient *U3*

The negative predictive value (NPV), positive predictive value (PPV), sensitivity, and specificity for *U3* across the simulation conditions are given in Figures B46, B47, B48, and B49, respectively. Additionally, Table A17 contains the NPV, PPV, sensitivity, and specificity aggregated across simulation conditions (dimensions, number of response categories, and test length).

While all the ANOVA terms were significant due to large sample size, the effect sizes provided some interesting results. Sensitivity and NPV showed large effects from the number of response options, test length, and a medium sized effect for the dimension by response option interaction. Specificity showed large effects from the number response options and test length. Additionally, specificity showed medium effects from both the dimension by number of response options and the number of response options by test length interactions. PPV showed medium or large effects on all terms except the main effect of dimension and the interaction between dimension and test length.

For NPV and sensitivity, the two-way interaction between response options and dimensionality is shown in the differences in accuracy across dimensions for each response option condition. The largest difference in outcomes for the number of response option condition was seen at 1-dimension and the smallest differences seen at 3- and 4-dimensions. For specificity, the dimension by response option interaction was similar to NPV and sensitivity. The test length by response option interaction is more difficult to visualize. However, there was an increase in accuracy in the medium length test condition compared to a short test condition. Though, the rate of increase between lengths was different depending on the response option condition. Four, five, and seven response options tended to improve by the same amount between the short test and medium test conditions. However, six response options showed more of an improvement in accuracy compared to the other response options between the two test lengths. Finally, PPV showed a meaningful three-way interaction. The three-way interaction is easily visualized in Figure B47. On medium tests, there is a large increase in the accuracy of six response options in the 1- and 4-dimension conditions when compared to short tests. This increase in accuracy was larger than the increases seen in the same conditions for other response option conditions. Additionally, seven response options showed a similar unique improvement in PPV for the 3-dimension condition between short and medium tests. This improvement in PPV was not shown in the other response option conditions.

Overall

Across all conditions, the number of normed Guttman errors outperformed *U3* and H_i^T. In aggregate, the number of Guttman errors showed .92 negative predictive value (NPV), .52 positive predictive value (PPV), 26.42% sensitivity, and 97.27% specificity.

Coefficient *U3* had the second-best performance, showing .92 NPV, .46 PPV, 23.87% sensitivity, and 96.96% specificity. Coefficient H_i^T was the worst performer in terms of classifying extreme response style; H_i^T showed .90 NPV, .10 PPV, 4.85% sensitivity, and 94.91% specificity.

Social Desirability Responding

Guttman Errors

The negative predictive value (NPV), positive predictive value (PPV), sensitivity, and specificity for the normed number of Guttman errors across the simulation conditions are given in Figures B50, B51, B52, and B53, respectively. Additionally, Table A18 contains the NPV, PPV, sensitivity, and specificity aggregated across simulation conditions (dimensions, number of response categories, and test length).

With such large sample sizes, all terms in the ANOVA for all outcomes were statistically significant. However, the effect sizes provided more useful information. For sensitivity and NPV, all effects sizes were medium or large for all terms except for the dimension by test length and number of response options by test length interactions. For specificity and PPV, all terms showed medium or higher effect sizes except the dimension by length interaction.

For NPV and sensitivity, the three-way interaction is easy to visualize from both Figures B50 and B52. Specifically, there are small differences between medium and short tests for all response option conditions at 1- and 2-dimensions. However, the seven-response option condition showed a large increase in accuracy on medium tests at 3- and 4-dimensions over and above the other response option conditions. While PPV and specificity showed different effect sizes for other terms, the three-way interaction is

similar: seven response options showed a large increase in accuracy at 3- and 4-dimensions on medium tests not seen in other response option conditions.

Coefficient H_i^T

The negative predictive value (NPV), positive predictive value (PPV), sensitivity, and specificity for H_i^T across the simulation conditions are given in Figures B54, B55, B56, and B57, respectively. Additionally, Table A19 contains the NPV, PPV, sensitivity, and specificity aggregated across simulation conditions (dimensions, number of response categories, and test length).

The effect sizes were used to determine the factors with medium or high impact on the outcomes given the overpowered ANOVAs. For sensitivity, all terms showed medium or large effects except the main effect of dimensionality. For specificity, all terms showed medium or large effects except for the three-way interaction, the main effect of test length, and the main effect of dimensionality. For NPV and PPV, all terms showed medium or large effects.

While NPV, PPV, and sensitivity showed different magnitudes of the effect, the three-way interaction had a similar interpretation in all three outcomes. For all three outcomes, the six and seven response option conditions show the biggest increase in accuracy between short and medium tests. Additionally, this increase in accuracy is much larger in the 2- and 4-dimension conditions compared to the 1- and 3-dimension conditions. For sensitivity, all three two-way interactions must be interpreted. For the dimensionality by response option interaction, there was an increase in accuracy between subsequent dimensions (i.e., 1- to 2-dimensions) for every response option condition except seven response options. However, from 3- to 4-dimension, seven response options

finally increased in accuracy rather than decreasing. For the test length by response option interaction, there was an increase in accuracy between short and medium length tests for all response options except four. Four response options showed an aggregate decrease in accuracy on medium tests compared to short tests. Finally, for the dimension by test length interaction, the aggregate accuracy on dimensions 1, 3, and 4, increased slightly between short and medium length tests. However, dimension 2 showed a very large increase in accuracy on medium tests compared to short tests. Dimension 2 showed the worst aggregate accuracy on short tests, but the best aggregate accuracy on medium tests.

Coefficient *U3*

The negative predictive value (NPV), positive predictive value (PPV), sensitivity, and specificity for *U3* across the simulation conditions are given in Figures B58, B59, B60, and B61, respectively. Additionally, Table A20 contains the NPV, PPV, sensitivity, and specificity aggregated across simulation conditions (dimensions, number of response categories, and test length).

The effect sizes provided more useful information regarding the importance of the model terms given the overpowered ANOVAs. For sensitivity, the main effect of response options and the interaction between dimensionality and response options showed medium effects. Specificity showed no medium or large effects. NPV and PPV both showed medium and large effects for all terms except the two-way interactions between dimensionality and test length and between response options and test length.

In terms of sensitivity, the two-way interaction between dimensionality and response options is difficult to visualize. However, as dimensionality increases, seven

response options tended to show a much larger increase to sensitivity than the other response option conditions. The five-response option condition tended to show a decrease in accuracy as dimensionality increased, and the four-response option condition showed an increase in sensitivity until the increase from 3- to 4-dimensions, where it showed a noticeable decrease. For both NPV and PPV, the three-way interaction was apparent from looking at Figures B58 and B59. Specifically, seven response options showed a large increase to both NPV and PPV on medium tests compared to short tests in the 3- and 4-dimension conditions. This improvement in NPV and PPV was not seen in other response option conditions.

Overall

Across all conditions, coefficient H_i^T showed the best PPV, sensitivity, and specificity when compared to Guttman errors and *U3* (there are negligible difference in NPV between the three person-fit statistics) when identifying social desirability responding. Coefficient H_i^T showed: PPV = .18, sensitivity = 9.40%, and specificity = 95.41%. Guttman errors performed second best, with PPV = .14, sensitivity = 7.15%, and specificity = 95.15%. Coefficient *U3* showed the worst performance when identifying social desirability responding. Specifically, it showed: PPV = .12, sensitivity = 6.25%, and specificity = 94.99%.

Careless Responding

Guttman Errors

The negative predictive value (NPV), positive predictive value (PPV), sensitivity, and specificity for the normed number of Guttman errors across the simulation conditions are given in Figures B62, B63, B64, and B65, respectively. Additionally, Table A21

contains the NPV, PPV, sensitivity, and specificity aggregated across simulation conditions (dimensions, number of response categories, and test length).

The results of the ANOVAs were not immediately useful due to the large sample sizes making all terms significant. However, the effect sizes were obtained and were more useful in determining the impact of the ANOVA terms on the outcomes. For sensitivity and NPV, all terms were medium or large except for the two-way interaction between dimensionality and test length. For specificity and PPV, all terms were medium or large.

In terms of sensitivity and NPV, four, five, and six response options all showed an increase in outcomes between short and medium length tests at all dimensions. However, four and six response options showed particularly large increases in both outcomes at 3 dimensions. The seven-response option condition showed minor increases in both outcomes at 1- and 2-dimenions between short and medium tests but showed larger increases in outcomes at 3- and 4-dimensions. For specificity and PPV, there is a similar relationship between short and medium length tests and seven response options. Specifically, seven response options showed a minor increase in outcomes at 1- and 2-dimensions between short and medium tests, and larger increases in outcomes at 3- and 4-dimensions. Six response options showed large increases between short and medium tests at the 1- and 3-dimension conditions compared to the other dimensionality conditions.

Coefficient H_i^T

The negative predictive value (NPV), positive predictive value (PPV), sensitivity, and specificity for H_i^T across the simulation conditions are given in Figures B66, B67,

B68, and B69, respectively. Additionally, Table A22 contains the NPV, PPV, sensitivity, and specificity aggregated across simulation conditions (dimensions, number of response categories, and test length).

Since all ANOVA terms were significant, the effect sizes were used to determine the impact of the factors on the outcomes. For sensitivity, all terms showed medium or large effects except for the interaction between dimensionality and test length. For specificity, all terms had medium or large effect sizes except for the three-way interaction and the two-way interaction between dimensionality and test length. For both NPV and PPV, all terms were medium or large.

For sensitivity, the three-way interaction was most noticeable in the four and five response option conditions. Both of these conditions showed an increase in sensitivity from short to medium tests. However, the five-response option condition showed a much larger increase in sensitivity at the 1-, 2-, and 3-dimension conditions compared to the four-response option condition.

For specificity, the two-way interactions between response options and test length and dimensionality by response options are the highest-level terms that showed medium effect sizes. For the dimension by response option interaction, all response options show a decrease in aggregate specificity between 1- and 2-dimensions (with six response options showing the largest decrease). Between 2- and 3- dimensions there is an increase in specificity for four and five response options and minor differences for six and seven response options. Between dimensions 3 and 4, all response options increase except for five response options, which shows a minor decrease in specificity. The two-way interaction between response options and test length showed response options four, five,

and six increasing similarly between short and medium length tests. However, the seven-response option condition showed an aggregate decrease in specificity between short and medium length tests.

For NPV and PPV, the three-way interaction was the highest-order medium or large effect. The three-way interaction was most noticeable when examining the differences between the four and five response option conditions across dimensions and test lengths. Specifically, on dimension 2, there was an increase in outcomes between short and medium tests for five response options, but little or no increase for four response options. Five response options also showed a larger increase than four response options between short and medium length tests on dimensions 1 and 3.

Coefficient *U3*

The negative predictive value (NPV), positive predictive value (PPV), sensitivity, and specificity for *U3* across the simulation conditions are given in Figures B70, B71, B72, and B73, respectively. Additionally, Table A23 contains the NPV, PPV, sensitivity, and specificity aggregated across simulation conditions (dimensions, number of response categories, and test length).

All ANOVA terms were statistically significant due to large sample sizes, so the effect sizes were used to determine the factors with meaningful impact on the outcomes. Sensitivity showed all effect sizes as large or medium except the two-way interaction between dimensionality and test length. Specificity, PPV, and NPV had all terms showing large or medium effects.

While all outcomes showed different effect sizes for different factors, the interpretation of the three-way interaction is the same across outcomes. The three-way

interaction is most noticeable when looking at four and five response options across test lengths at different dimensionalities. Specifically, on short tests, four and five response options are somewhat differentiated in terms of outcomes until 3- and 4-dimensions, where their outcomes are more similar. On medium length tests, four and five response options are similar at 1-, 2-, and 4-dimensions.

Overall

Across all conditions, the normed number of Guttman errors provided the highest classification accuracy in terms of PPV, NPV, sensitivity, and specificity for identifying careless responding. Specifically, Guttman errors showed: NPV = .92, PPV = .56, sensitivity = 26.68%, and specificity = 97.54%. Coefficient *U3* showed the next best performance, with: NPV = .92, PPV = .50, sensitivity = 25.30%, and specificity = 97.18%. Finally, H_i^T showed the worst performance, with: NPV = .91, PPV = .25, sensitivity = 12.72%, and specificity = 95.77%.

Aggregated Results

In aggregate, a few patterns started to emerge among the results. Tables A24, A25, and A26 present the results aggregated results across simulation conditions for each person-fit statistic and aberrant response pattern. Across response styles there is little difference (never greater than ± 0.02) between the negative predictive values (NPV) and specificity for any of the person-fit statistics across any of the aberrant response patterns. These findings suggest that all person-fit statistics are capable of correctly identifying individuals who are not aberrant responders with a high degree of accuracy (specificity). Additionally, once the person-fit statistics identify a respondent as non-aberrant, there is a high probability that they are indeed a non-aberrant responder (NPV).

However, the person-fit statistics begin to differentiate themselves when looking at the sensitivities and positive predictive values (PPV). Sensitivity is identical to statistical power, and none of the person-fit statistics obtained the 0.80 convention for statistical power for any aberrant response pattern. Regardless, Guttman errors showed the highest aggregate sensitivity (.13), followed by U3 (.12), and finally by H_{i}^{T} (.09). Examining the results by aberrancy, Guttman errors showed the highest sensitivity when identifying disacquiescence (.13), extreme responding (.27), and careless responding (.29). Coefficient H_i^T showed the highest sensitivity when identifying midpoint (.13) and careless responding (.13). Coefficient *U3* showed the highest sensitivity when identifying disacquiescence (.12), extreme responding (.25), and careless responding (.25). However, the PPV values are more damning. The only aggregate conditions where those identified as aberrant had greater than a .50 probability to truly be aberrant is when using Guttman errors to identify extreme and careless responding.

While many of these interpretations changed based on the condition (i.e., changed by length, dimensionality, and number of response options) they were being applied to, the aggregate results are still informative about the overall performance of the person-fit statistics. Tables A6 - A23 are provided to explore the results when they are aggregated by the simulation conditions.

Figures B74 – B79 show the results aggregated over aberrant response patterns. Across all number of response option conditions, four and five response options showed the highest outcomes for Guttman errors, followed by six response options, with seven response options showing the lowest outcomes. However, this depends on dimensionality and test length, and the difference between four and five response options increased or

decreased based on these factors. Both five and seven response options tended to increase as dimensionality increased, while four and six response options tended to decrease as dimensionality increased. Finally, the medium test length condition almost always showed higher outcomes than the short test length condition.

Coefficient H_i^T almost always showed increased outcomes between short and medium length tests on all outcomes. Additionally, coefficient H_i^T showed an interesting trend in terms of dimensionality. Specifically, 1-dimension showed the highest outcomes which was followed by a large drop in outcomes at 2-dimensions. There was an increase in outcomes at the 3- and 4-dimension conditions, but the increases never reached the peak in outcome values at 1-dimension. On medium length tests, the seven-response option condition is fairly consistent across dimensionalities.

Coefficient *U3* showed similar trends as Guttman errors. Outcomes using coefficient *U3* almost always benefited from increased test length, having four or five response options, and tended to perform the best at 1- and 2-dimension conditions (though, this was dependent on the three-way interaction and is not true in all cases).

With the results provided, the study's hypotheses can be resolved. **Hypothesis 1** stated that H_i^T will show greater sensitivity, specificity, positive predictive values, and negative predictive values when compared to *U3* and the number of Guttman errors. Interestingly, hypothesis 1 was rejected in most cases. As discussed, in aggregate, H_i^T only outperformed U3 and Guttman errors in acquiescence responding, midpoint responding, and socially desirable responding.

Hypothesis 2 stated that within a person-fit statistic, the sensitivity, specificity, positive predictive values, and negative predictive values will be similar across

aberrancies. Hypothesis 2 was also rejected. As seen in the aggregate tables (Tables A24 – A26) there is much variability in the sensitivity and PPV of the person-fit statistics depending on the type of aberrant responding. However, the specificity and the NPV do remain relatively stable across aberrancies.

Hypothesis 3 stated that all person-fit statistics will show increased sensitivity, specificity, positive predictive values, and negative predictive values as the number of response options increased. Hypothesis 3 is rejected due to the many three-way interactions between all simulation conditions. At the very least, this study cannot proffer strong evidence to suggest that increasing the number of response options will increase the accuracy of person-fit statistics across the conditions. Hypothesis 3 had a few results that supported it (e.g., H_{i}^{T} when identifying acquiescence or socially desirable responding). However, the number of response options often showed a pattern that favored either an odd or even number of response options, fewer response options (i.e., 4 and 5 response options outperformed 6 and 7), or the response options showed the opposite of the hypothesized pattern (i.e., outcome values decrease as the number of response options increased, as seen for careless responding).

Hypothesis 4 stated that the person-fit statistics will show decreased sensitivity, specificity, positive predictive values, and negative predictive values as the dimensionality of data increased. Similar to hypothesis 3, hypothesis 4 was rejected due to the many three-way interactions between the simulation conditions. Hypothesis 4 had the least corroborating evidence supporting it; only one condition supported hypothesis 4: The number of Guttman errors when identifying socially desirable responding. Otherwise, the impact of dimensionality was quite variable by type of aberrancy.

Hypothesis 5 stated that the person-fit statistics will show increased sensitivity, specificity, positive predictive values, and negative predictive values as test length increased. While hypothesis 5 is the most difficult of the hypotheses to reject, it was simply not true in all cases due to the three-way interactions. Both the accuracy of *U3* and the number of Guttman errors failed to improve as test length increased for disacquiescence responding and midpoint responding, respectively.

CHAPTER V: DISCUSSION

Summary of Findings

As a preface, it must be emphasized that all aggregated results should be considered through the lens of the practically meaningful two- and three-way interactions. These interactions were found for most outcomes across aberrant response patterns. Additionally, the person-fit statistics operated differently across the various aberrant response patterns, effectively resulting in four-way interactions (i.e., aberrancy by number of response options by dimensionality by test length). Four-way interactions make the interpretation of results very complex. However, there were still several patterns that emerged in the outcome variables.

Impact of the Simulation Conditions

In aggregate, there was little difference between the ability of the nonparametric person-fit statistics to correctly identify individuals not responding aberrantly; all of the person-fit statistics showed expected (or higher) specificity and high negative predictive values (NPV) across most aberrancies. However, the simulation conditions still had a small impact on NPV and specificity estimates. While NPV and specificity estimates were fairly stable across person-fit statistics, there was differentiation when examining PPV and sensitivity estimates. Even with said differentiation, considering the traditional rules-of-thumb for adequate statistical power ($1 - \beta = 0.80$), none of the person-fit statistics showed acceptable sensitivity estimates or high PPV estimates. Indeed, only Guttman errors and *U3* showed any PPV estimates greater than 0.50. However, the simulation conditions still had an impact on the PPV and sensitivity estimates.

Impact on Guttman Errors

For Guttman errors, NPV, PPV, specificity, and sensitivity estimates benefitted from increased test length, almost regardless of the other simulation factors. Midpoint responding was the only aberrancy where Guttman errors did not show increased accuracy due to increased test length. While there were minor differences in the sensitivity estimates for Guttman errors between test lengths, the estimates were so small that any differences were negligible. The overall observed increase in accuracy for Guttman errors as test length increased is not a surprising finding. Many studies have shown that the power of Guttman errors to detect aberrancy increases with test length in both dichotomous and polytomous response data (Emons, 2008; Karabatsos, 2003; Meijer, 1994; St-Onge et al., 2011).

When examining the number of response options, four and five response options resulted in the highest outcomes across most aberrancies. Seven response options resulted in the lowest outcomes across most aberrancies (this finding is particularly clear in Figures B74 and B75). The six-response option condition generally performed worse than four and five response options but better than the seven-response option condition. However, when examining specific aberrancies, seven response options showed the highest performance when identifying social desirability responding. Compared to the next highest performing number of response categories (five), seven response categories showed a 2.26% increase in sensitivity with small increases for NPV, PPV, and specificity.

In terms of dimensionality, all outcomes tended to perform worse as dimensionality increased (when aggregated across aberrancies). The decrease in accuracy

due to dimensionality was particularly noticeable for four, five, and six response options. The only exception was the seven-response option condition, which tended to perform better as dimensionality increased. The impact of dimensionality also heavily depended on the type of aberrancy being identified. Acquiescence and socially desirable responding tended to show increased accuracy as dimensionality increased; disacquiescence and careless responding tended to show decreased accuracy as dimensionality increased; and midpoint and extreme responding showed unique patterns. Specifically, midpoint responding had the highest accuracy at two and three dimensions, while extreme responding had the highest accuracy at one and four dimensions.

In terms of identifying aberrancies, Guttman errors showed the highest sensitivity when identifying extreme (sensitivity = 0.27) and careless responses (sensitivity = 0. 29). The number of Guttman errors showed mediocre sensitivity when identifying disacquiescence responding (sensitivity = 0.13) and performed poorly when identifying acquiescence (sensitivity = 0.05), midpoint (sensitivity < 0.01), and socially desirable responding (sensitivity = 0.08). Additionally, Guttman errors showed higher than expected specificity (0.95 based on the chosen cutoff) for disacquiescence (specificity = 0.96), extreme (specificity = 0.97), and careless responding (specificity = 0.98). However, Guttman errors also showed lower than expected specificity for midpoint responding (specificity = 0.94).

Impact on Coefficient $H_{\mathbf{i}}^{T}$

For coefficient H_{i}^{T}, NPV, PPV, sensitivity, and specificity estimates tended to improve in the medium-test length condition compared to the short-test length condition, almost regardless of other factors. The only exception to improved accuracy with

increased test length was the four-response option condition at 3- and 4-dimensions. When aggregated across aberrant response styles, the four-response option condition tended to perform better in the short-test condition at higher dimensionalities. The general trend of increased test length improving the accuracy of coefficient H_i^T is not surprising, as several studies have shown that coefficient H_i^T increases in power as test length increases (Karabatsos, 2003; St-Onge et al., 2011). However, this effect had not yet been demonstrated in polytomous response data.

When aggregated across aberrant response patterns, the impact of the number of response options on the outcome variables was heavily dependent on dimensionality and test length. Four or seven response options tended to show the best performance on short tests, while six or seven response options tended to show the best performance on medium tests. Additionally, the impact of the number of response options on the accuracy of coefficient H_i^T was heavily dependent on the type of aberrancy being identified. In general, coefficient H_i^T tended to show improved accuracy as the number of response options increased when identifying acquiescence, midpoint, and socially desirable responding. Conversely, coefficient H_i^T generally showed a decrease in accuracy when identifying disacquiescence, extreme, and careless responding.

In terms of dimensionality, H_i^T tended to show the highest sensitivity, specificity, NPV, and PPV outcomes in the 1- or 4-dimension conditions. However, this finding was heavily dependent on the number of response options and test length. On short tests, the 1-dimension condition showed the highest performance for six and seven response options, while the 4-dimension condition showed the highest outcomes on four and five response options. On medium tests, four, five, and six response options showed the

highest outcomes at 1-dimension. Conversely, seven response options showed the highest outcomes at 4-dimensions for medium-length tests. While most response option conditions were variable across dimensions, the seven-response option condition remained relatively stable across dimensions on medium-length tests.

When aggregated over all other conditions, coefficient H_i^T showed mediocre accuracy when identifying midpoint (sensitivity = 0.13) and careless responding (sensitivity = 0.13) and performed poorly when identifying acquiescence (sensitivity = .08), disacquiescence (sensitivity = 0.04), extreme (sensitivity = 0.05), and socially desirable responding (sensitivity = 0.09). However, coefficient H_i^T reached the expected specificity (.95) for every aberrant response pattern. Additionally, it showed higher than expected specificity estimates for midpoint (specificity = .96) and careless responding (specificity = .96).

Impact on Coefficient *U3*

The overall patterns for coefficient *U3* are similar to the patterns observed for the number of Guttman errors. Increasing test length tended to result in improved accuracy for *U3*. However, there was no noticeable increase in accuracy between short and medium-length tests for midpoint responding. Additionally, there was a slight decrease in sensitivity and specificity at the medium-test length condition compared to the short-test condition for disacquiescence responding. The improvement in the accuracy of *U3* when identifying most aberrant response patterns as test length increases is not a surprising finding. Several studies have shown that increasing test length improves the power of *U3* to detect a variety of aberrant response patterns in both dichotomous and polytomous response data (Emons, 2008, Karabatsos, 2003; St-Onge et al., 2011).

In general, four and five response options showed higher accuracy when aggregating results across aberrant response patterns. Seven response options showed the worst performance in all conditions, with six response options generally performing worse than four and five response options but better than seven response options. However, the effect of the number of response options was heavily impacted by the type of aberrancy being identified. Generally, disacquiescence and careless responding showed decreased accuracy as the number of response options increased. Conversely, social desirability responding showed increased accuracy as the number of response options increased. Acquiescence, midpoint, and extreme responding all showed unique patterns associated with the number of response options. Specifically, acquiescence showed the highest outcomes at five response options with minor differences between the other response options; midpoint responding showed the best performance at four and seven response options (but performed poorly overall); and extreme responding showed the best performance with five response options and the worst performance with seven response options.

In terms of dimensionality, the accuracy of *U3* generally decreased as dimensionality increased, but this trend was heavily impacted by the type of aberrancy being identified. Specifically, coefficient *U3* showed decreased accuracy as dimensionality increased when identifying acquiescence, midpoint, extreme, and carless responding. Conversely, coefficient *U3* showed a general increase to accuracy as dimensionality increased when identifying disacquiescence and social desirability responding, though the increases were dependent on other factors.

In terms of the ability of *U3* to identify specific aberrant response patterns, there is little that differentiates *U3* from the number of Guttman errors. Coefficient *U3* showed the highest sensitivity when identifying extreme (sensitivity = 0.25) and careless responding (sensitivity = 0.25); mediocre sensitivity when identifying disacquiescence responding (sensitivity = 0.12); and poor sensitivity when identifying acquiescence (sensitivity = 0.05), midpoint (sensitivity < 0.01), and socially desirable responding (sensitivity = 0.07). Additionally, *U3* showed higher than expected specificity (0.95) when identifying disacquiescence (specificity = 0.96), extreme (specificity = 0.97), and careless responding. (specificity = 0.97). However, coefficient *U3* showed lower than expected specificity when identifying midpoint responding (specificity = 0.94).

Comparing Aberrant Response Patterns and Person-fit Statistics

Interestingly, Guttman errors and *U3* showed remarkably similar results throughout the study. Across conditions, both person-fit statistics showed higher outcome values than coefficient H_i^T (e.g., both had comparably high sensitivity when identifying careless responding). Additionally, they showed similar outcome patterns across most other simulation conditions. However, Guttman errors slightly and consistently outperformed *U3* in nearly every condition. While the similarity between the statistics is not surprising, as both *U3* and Guttman errors are measures of nonconformity, the advantage in performance held by Guttman errors is more difficult to explain. In fact, Emons (2008) showed that *U3* generally outperformed Guttman errors when used on items with two response options when identifying careless responding, but that Guttman errors generally outperformed *U3* when used on items with four response options. However, the differences between the two statistics in both conditions were small. Emons

suggested that the difference in performance can be explained by the use of deviance scores in the calculation of *U3*. Deviance scores result in U3 showing higher power when there are stark changes in a response set (see Emons, 2008, Table 5). Conversely, since Guttman errors are based on a simple normed count, they are capable of identifying subtler deviations in response sets (Emons, 2008).

Coefficient H_i^T differentiated itself from Guttman errors and *U3*, while not performing as well overall. Coefficient H_i^T was the only person-fit statistic that showed any ability to detect midpoint responding. However, midpoint responding was the only aberrant response pattern where coefficient H_i^T had higher outcomes than *U3* or Guttman errors. Interestingly, the number of response options and dimensionality (i.e., the interaction effects) had a more noticeable impact on coefficient H_i^T than on *U3* or Guttman errors (clearly seen in Figures B76 - B77). The larger impact of the interaction effects possibly results from how H_i^T is calculated. In contrast to *U3* and Guttman errors, H_i^T is a measure of conformity between a response vector and the other response vectors in a sample. In effect, it is a correlation between an individual's response vector and a group's summarized response vector (Karabatsos, 2003; St-Onge et al., 2001). Given this idiosyncrasy, H_i^T is more sensitive to individual changes in item responses across a data set (Karabatsos, 2003).

Across all conditions, Guttman errors and *U3* showed low sensitivity for midpoint, socially desirable, and acquiescence response behaviors. Coefficient H_i^T showed low sensitivity when identifying acquiescence, disacquiescence, extreme, and socially desirable response behaviors. Since the person-fit statistics performed well in terms of specificity, the most likely explanation for this poor performance is that the

nonparametric person-fit statistics lacked statistical power within each simulated dataset. The lack of statistical power is likely caused by a combination of two issues. 1) First, this dissertation chose a conservative cutoff that limited type I errors. While this cutoff was chosen to provide an easy point of comparison for the statistics, a conservative cutoff will reduce statistical power (Bilder & Loughin, 2015). 2) Second, even with the large sample sizes used in this dissertation, an aberrant responder was still modeled as a relatively rare occurrence (only 100 out of every 1000 responders were modeled as aberrant). Rare outcomes can affect the statistical power of studies and are often considered when conducting power analyses in medical and biological research (Buderer, 1996).

Discussion and Implications

Nonparametric person-fit statistics have been used with varying levels of success to identify aberrant response patterns in dichotomous response data, often showing equivalent or better performance than their parametric counterparts (Dimitrov & Smith, 2006; Emons, 2008; Karabatsos, 2003; Meijer, 1994; Niessen et al., 2016, St-Onge et al., 2011; Tendeiro & Meijer, 2014). However, not much research has investigated the applicability of nonparametric person-fit statistics to polytomous response data. This dearth of research is largely due to the lack of easily available generalizations of many person-fit statistics. However, when studies investigated the use of nonparametric person-fit statistics with polytomous response data, findings have been mixed.

Emons (2008) simulated unidimensional polytomous response data for careless and extreme responding. He examined the impact of test length, the number of response options, item discrimination, and the proportion of aberrant responses in a response vector on the accuracy of the number of Guttman errors and *U3* (among others). Emons

found that Guttman errors tended to slightly outperform *U3* in most conditions for careless responding. For extreme responding, Emons also found that Guttman errors slightly outperformed *U3*; additionally, Emons found that both nonparametric person-fit statistics (Guttman errors and *U3*) performed similarly to a parametric person-fit statistic. This dissertation showed a similar relationship between Guttman errors and *U3*. Namely, Guttman errors, with few exceptions, slightly outperformed *U3* in most conditions.

Additionally, Emons (2008) found that *U3* and Guttman errors were able identify careless responding at a higher rate than extreme responding. In contrast, this dissertation found few differences between the aggregate identification rates of careless and extreme responding for Guttman errors or *U3*. However, it should be noted that Emons only examined two and four response categories. This dissertation found a decrease in the sensitivity and PPV of Guttman errors and *U3* at six and seven response categories, but the four- and five-response category conditions show PPV and sensitivity estimates much closer to the results found by Emons. Similarly, the seven-response category condition also resulted in lower PPV and sensitivity estimates for Guttman errors and *U3* when identifying extreme responding. When looking at the response category conditions with the fewest options (i.e., four and five), the sensitivity and PPV estimates are closer to those found by Emons.

Therefore, both studies support the conclusion that *U3* and Guttman errors perform better with fewer response options when identifying careless responding compared to extreme responding. However, as the number of response options increases, the difference in identification rate of *U3* and Guttman errors between extreme and careless responding decreases. Specifically, there are only small differences in the

performance of the person-fit statistics at six response categories, and at seven response categories both person-fit statistics show higher sensitivity when identifying extreme response behavior compared to careless response behavior. This finding is likely a result of how careless responding was simulated: as random responding. As the number of response options increases, the number of possible unique response patterns also increases (Cox, 1980). As the number of unique response patterns increases, it becomes more difficult to identify Guttman errors as there is higher likelihood that a response not representing a Guttman error will be chosen at random. Additionally, for *U3*, as the number of unique response patterns increases, the probabilities of the most and least possible deviant response patterns (*Pmin* and *Pmax* from Equation 15) become more extreme (van der Flier, 1982).

This dissertation showed the number of response categories had a large impact on person-fit statistic accuracy for most aberrancies. In contrast, Emons (2008) found only a small effect of the number of response categories on careless responding and a noticeable impact on extreme responding when using *U3* and Guttman errors. However, Emons limited their simulation to two and four response categories. It is possible that Emons findings are due to restriction of range, as this dissertation did not start to see a large impact on careless or extreme responding until investigating the six-response category condition and beyond. Additionally, the shift between four and five or five and six response categories is where the number of response options has the most impact in other aberrant response patterns (if there was a noticeable impact of response categories on the outcomes).

Emons (2008) also found a noticeable increase in accuracy between two and four response categories when using *U3* and Guttman errors to identify extreme responding. Based on this finding, Emons hypothesized that having items with five or more response categories would improve the accuracy of *U3* and normed Guttman errors. This dissertation did find an increase in the accuracy of *U3* and normed Guttman errors when identifying extreme responses between four and five response categories, but the accuracy began to decrease at six and seven response categories. Based on findings from both studies, it seems that there might be a benefit to the accuracy of U3 and Guttman errors when identifying extreme responses up until five categories, after which there are deleterious effects. While there are interactions between response categories, test length, and dimensionality to consider, five response categories tended to offer the best performance for *U3* and Guttman errors across conditions, being outperformed slightly by four and six response categories in specific conditions (see Figures B74 – B79).

In terms of test length, this study corroborates the findings from Emons (2008). In nearly all conditions, Emons found that increasing test length (from 12- to 24-items) improved the detection rates of *U3* and Guttman errors when identifying careless and extreme responding. Similarly, this study found that increasing test length (from 12- to 36-items) nearly always improved the detections rates of *U3*, Guttman errors, and H^T_i. This finding also supports the consensus on test length found in studies focused on dichotomous response data (Dimitrov & Smith, 2006; Karabatsos, 2003; Meijer, 1994; Niessen et al., 2016, St-Onge et al., 2011; Tendeiro & Meijer, 2014).

Beck et al. (2019) applied *U3*, the normed number of Guttman errors and H^T_i, to a multidimensional, polytomous data set containing a measure of careless responding (a

single instructed response item). The person-fit statistics were used to predict the outcome of the instructed response item, from which a Receiver Operating Curve (ROC) was created. Beck et al. found that H_i^T provided the highest area under the ROC, suggesting that it was outperforming the other statistics at predicting the outcome of the instructed response item. However, when using the area under the ROC to find an empirical cutoff, Beck et al. found that using H_i^T to remove individuals from a data set provided mixed results in terms of improvement to measurement model fit. Additionally, Beck et al., found that Guttman errors and *U3* performed poorly in terms of improvement to model fit, and in their ability to predict the outcome of the instructed response item (shown by low area under the curve based on the ROC).

The data used by Beck et al. (2019) were from a 34-item, self-report survey with four factors. Additionally, the items on the survey were Likert-type and had five response categories. According to the findings from this dissertation, a survey with these characteristics should result in *U3* and Guttman errors showing higher sensitivity to careless responding than H_i^T (i.e., sensitivity [Guttman] = 47.43%; sensitivity [*U3*] = 44.66%; sensitivity [H_i^T] = 22.79%; see Figures B62 – B73). This finding is contrary to what Beck et al. found in their application of these person-fit statistics to a real data set. While these findings and the findings of Beck et al. do appear contradictory, they can be reconciled.

The sensitivity and specificity of a classification method is a spectrum, and sensitivity and specificity are inversely proportional to each other. Decreasing the specificity of a classification method will increase its sensitivity and vice versa (Bilder & Loughin, 2015). To allow for direct comparisons, this dissertation chose a conservative

cutoff for all person-fit statistics. Coefficient H_i^T would have shown higher sensitivity for identifying (at least) careless responders if a more liberal cutoff was chosen. In fact, using person-fit statistics with different cutoffs creates the possibility of an effective two-stage approach to identifying careless responding (and other aberrant response patterns). Step one would be the application of a person-fit statistic with high sensitivity as a screening test for the aberrancy. Step two would be the application of a person-fit statistic with high specificity to identify cases from step one that represent likely Type II errors. A two-step approach as described could be a practical and effective method for identifying aberrancy, but more research would be needed to determine if it was appropriate.

In summary, the findings of the current study are more comparable to Emons (2008) than Beck et al. (2019). Similar to the findings in Emons, Guttman errors and *U3* performed well when identifying extreme and careless responding, with Guttman errors slightly outperforming *U3*. This result is not surprising, as the method to simulate both aberrant response patterns in both studies are identical. However, it does provide replicability of the results found by Emons. Beck et al. found that H_i^T provided the best predictability of the careless response indicator, and that Guttman errors and *U3* performed poorly at the same task. This dissertation showed the opposite results: Guttman errors and *U3* should be outperforming H_i^T when identifying careless responses under the same conditions. The consensus in the extant literature is that H_i^T has great potential at identifying aberrant response patterns in dichotomous response data (Dimitrov & Smith, 2006; Karabatsos, 2003; Meijer, 1994; Niessen et al., 2016, St-Onge et al., 2011; Tendeiro & Meijer, 2014). Even research that is critical regarding the

application of H_i^T has suggested that it performs similarly to parametric person-fit statistics when identifying aberrant responders in dichotomous data (Sinharay, 2017).

The difference in findings between Beck et al. (2019), this dissertation, and the extant literature can possibly be due to the conservative cutoff chosen by this dissertation. Beck et al. examined three different cutoffs derived from ROC results, and the extant literature has employed a variety of methods to determine cutoffs for these statistics. Finding optimal cutoffs for these person-fit statistics is an important step for making them practical. However, differences in cutoff methods do not explain why Guttman errors and *U3* showed poor performance in Beck et al. It is possible that the instructed response item used in Beck et al. was not operating as a valid measure of careless responding. While instructed response items work in theory, it is difficult to judge their efficacy in practice. The mechanisms behind careless responding are not well defined, as such it is difficult to determine if any *a priori* measure (such as instructed response items) are reliable and valid for the purposes of identifying said behavior.

It is also possible that the addition of response categories beyond two, or dimensions beyond one, make it difficult for H_i^T to identify careless responding. There is a fairly large decrease in the sensitivity and PPV of H_i^T between 1-dimension and 2-dimensions for four and five response options (Figure B68). This decrease suggests that as dimensionality and response options increase, they have a deleterious impact on the ability of H_i^T to identify careless responding. Finally, it is also possible that the simulation of careless responding in this dissertation was simply unrealistic. This dissertation simulated careless responding as random responses across the response scale. However, given that H_i^T has been shown to predict careless responding accurately in other studies, it

is possible that random responding does not capture all the intricacies of real careless responding. Since Beck et al. employed real data, the difference between the two sets of results could simply be the fact that random responding is not the same as careless responding in real data.

Based on findings from this dissertation and the implication of these findings within the extant literature, a few recommendations can be made. First, nonparametric person-fit statistics perform better on longer tests. This result has been replicated many times in dichotomous response data and in studies investigating polytomous response data. There is little evidence to the contrary. Therefore, it can be stated with a degree of confidence that nonparametric person-fit statistics will more accurately identify aberrant responses on longer tests and/or surveys. Additionally, nonparametric person-fit statistics should be applied to shorter tests or surveys only when the reduction in accuracy is an acceptable trade-off for identifying aberrancy.

Second, the violation of the unidimensionality assumption does impact the accuracy of nonparametric person-fit statistics, though the impact is heavily dependent on other factors. Unfortunately, the impact was not consistent across aberrancy or person-fit statistics. Dimensionality showed a noticeable (PPV ± 0.1 between largest and smallest value) impact on Guttman errors when identifying disacquiescence responses; H_{i}^{T} when identifying midpoint responses; all person-fit statistics when identifying extreme responses; all person-fit statistics when identifying socially desirable responses; and on H_{i}^{T} when identifying careless responses. Interestingly, increasing dimensionality did not always decrease accuracy, as one would expect when violating a statistical assumption. In several conditions, increasing dimensionality improved the accuracy of the person-fit

statistic(s). For example, H_i^T with seven response categories often performed best in the 3- or 4-dimension conditions.

There were also several conditions where increasing the dimensionality had little to no impact on the accuracy of the person-fit statistic(s). As it stands, it seems that the nonparametric person-fit statistics can be robust to the violation of unidimensionality in certain situations. However, the exact nature of their robustness is still unclear. Regardless, the dimensionality of a survey in conjunction with the survey's other characteristics should be carefully considered when using nonparametric person-fit statistics. Different surveys may create different use-cases for different person-fit statistics. For example, when trying to identify disacquiescence on a medium length test, Guttman errors generally provide the highest accuracy. However, if the survey has 3- or 4-dimensions, *U3* actually shows higher sensitivity across response options.

Third, the number of response categories also impacts the ability of nonparametric person-fit statistics to detect aberrant response patterns in polytomous response data. Generally, there was a noticeable impact (PPV ± 0.1 between largest and smallest value) of response categories on Guttman errors and H_i^T when identifying acquiescence responses; all person-fit statistics when identifying disacquiescence responses; H_i^T when identifying midpoint responses; Guttman errors and *U3* when identifying extreme responses; Guttman errors and H_i^T when identifying socially desirable responses; and all person-fit statistics when identifying careless responses. It is difficult to arrive at a generalization based on these findings because of the inconsistency of the impact across person-fit statistics and type of aberrancy. However, the trend appears to be (at least on medium-length tests) that Guttman errors and *U3* show higher accuracy with fewer

response options (four and five), while H_i^T generally shows higher accuracy with more response options (six and seven). On the other hand, it could be noted that these general trends depend heavily on test length, type of aberrancy, and dimensionality (particularly for H_i^T). Regardless, it is safe to conclude that the number of response categories does impact the accuracy of nonparametric person-fit statistics.

Fourth, a single person-fit statistic cannot be used as a blanket test for all aberrant response patterns. Aberrant response patterns present differently, and evidence from this dissertation and the extant literature suggest that some aberrant response patterns are more difficult to detect than others (see Dimitrov & Smith, 2006; Emons, 2008; Karabatsos, 2003; Meijer, 1994; for a few examples). This information is not new, but it reiterates how crucial it is to use multiple methods and measures. While it is unfortunate that a single nonparametric person-fit statistic cannot serve as an omnibus test for aberrant responses, a combination of nonparametric person-fit statistics may be able to identify aberrant responses more accurately under optimal conditions. For example, Guttman errors performed quite well when identifying careless and extreme responses. Coefficient H_i^T was able to identify midpoint responding more accurately than *U3* or Guttman errors. Therefore, the combined usage of Guttman errors and H_i^T could provide a way to identify three types of aberrancies. While not a perfect solution, calculating multiple person-fit statistics to identify three or more aberrant response patterns is still fairly practical, particularly if the person-fit statistics were easily accessible in software packages.

Finally, while Guttman errors, in particular, are relatively adept at identifying extreme and midpoint responding, the sensitivity of these indices is still well below 50%.

In comparison, a general rule of thumb is that statistical power for any given test should be around 80% (Cohen, 1992b). Given the cutoff employed by this dissertation, the nonparametric person-fit statistics did not have the statistical power to reliably identify aberrancy. However, these nonparametric person-fit statistics would have better statistical power with a more liberal cutoff. Indeed, Meijer (2003) suggested that the ability of nonparametric person-fit statistics to detect aberrant response patterns might benefit from higher levels of alpha (i.e., greater than 0.05). Increasing alpha levels will increase Type I errors, so the difficulty arises in identifying a cutoff that provides acceptable sensitivity without too many Type I errors. This dissertation chose a cutoff under the assumption that data are valuable, and that researchers do not want to haphazardly discard quality data. Increasing the rate of Type I errors will increase the number of respondents being identified as aberrant. As seen in Beck et al. (2019), choosing a cutoff for these statistics that increased the rate of Type I errors led to a large proportion of the sample being identified as aberrant and removed. As such, modifying the Type I error rate needs to be done with care. On a more positive note, given an empirically bootstrapped cutoff, these nonparametric person-fit statistics are very accurate at identifying those individuals who are *not* engaging in aberrant responding. While contrary to how these statistics are generally applied in dichotomous data, these statistics could actually be used to confidently retain good responders with great accuracy rather than discard aberrant responders.

Limitations

No study is perfect, and this study is no exception. There are multiple limitations to this dissertation, and the findings should be considered in light of these limitations.

Perhaps the most glaring limitation is how the aberrant response patterns were simulated. The mechanisms of aberrant response patterns are not well studied. When the mechanisms of aberrant response patterns have been studied, the results and interpretations of said results have been met with disagreement and uncertainty. It is always difficult to simulate a complex human behavior, and complex human behaviors are unlikely to be recreated perfectly with any statistical model or simulation. So, while a literature-based and justifiable method was used to simulate all of the aberrant response behaviors, it is unlikely that the simulations captured all idiosyncrasies of these behaviors. In addition to these concerns, the simulations for socially desirable responding and careless responding were simplified due to practical reasons.

This study only examined the appropriateness of nonparametric person-fit statistics as indicators of aberrant responding. There is a myriad of other person-fit statistics in the extant literature (parametric and other nonparametric) designed for this purpose. As discussed earlier, many of the person-fit statistics have not been generalized for use in polytomous response data. While this dissertation investigated the accessible (i.e., generalized) nonparametric person-fit statistics, it would be informative to investigate the person-fit statistics that are less accessible.

This study used ANOVAs to identify differences between the simulation conditions. However, the large sample size rendered the ANOVAs effectively useless in terms of statistical significance; even using very strict *p*-level criteria ($p < .0001$) provided no benefit. While the effect size estimates were useful in identifying meaningful differences, the large sample sizes also created an issue for *post hoc* testing (i.e., Tukey's HSD). Due to this reason, the Tukey's HSD results were not presented, or discussed, in

this dissertation. *Post-hoc* testing should normally be conducted following an ANOVA test, but the absence of usable *post-hoc* results for this dissertation did not affect the interpretation of the results.

Finally, it would have been informative to apply the findings from this simulation study to a study utilizing real data. A study of this nature would allow for corroborating findings in a real-world setting. Additionally, having real-world data would create the opportunity to investigate the practical impacts of applying these person-fit statistics in a real-world setting.

Future Research

There is a dearth of research investigating person-fit statistics with polytomous response data, and there is much additional work to be done. This area is ripe for research, as nonparametric person-fit statistics could represent a practical and efficient method for improving validity in low-stakes testing contexts. Based on the results of this study, several future research ideas can be suggested.

First, a study should be conducted comparing the nonparametric person-fit statistics from this dissertation to parametric person-fit statistics in polytomous response data. The study would largely follow the method employed by this dissertation, with the addition of one or more parametric person-fit statistics. The multidimensional graded response model (MGRM) could be used to generate multidimensional, polytomous response data based on parameters from an extant survey. Then, aberrant response patterns would be generated, and person-fit statistics would be calculated on the resulting data. Using a bootstrapped empirical cutoff, the sensitivity, specificity, NPV, and PPV could be calculated. This proposed study could even investigate different empirical

cutoffs. Of particular interest would be the impact of dimensionality on the parametric person-fit statistics, and how their accuracy compares to the nonparametric person-fit statistics.

Second, more nonparametric person-fit statistics should be generalized to operate in polytomous data. While this undertaking is not a study, per se, it would have a large impact on person-fit research in the future. Generalizing nonparametric person-fit statistics under polytomous models would require a strong background in mathematics and statistics, and the ability to generate and solve mathematical proofs. Successfully generalizing person-fit statistics would likely result in a publishable paper. Additionally, once additional person-fit statistics were generalized, R packages would need to be created (or existing ones updated) to include the new generalizations.

Third, a study attempting to establish the sampling distributions of these nonparametric person-fit statistics would increase the likelihood that these statistics see widespread adoption. Establishing known and viable sampling distributions would mean that statistical hypothesis testing via calculating *p*-values would be possible for these person-fit statistics. Unfortunately, such an investigation would be complex and would not necessarily result in a publishable paper. One would need to investigate how well these person-fit statistics meet the assumptions and approximate various discrete probability distributions (e.g., Poisson, binomial, Bernoulli, etc.), or other relevant distributions. While this study would require a deep understanding of statistical and distributional research, it would have a huge impact on person-fit and aberrant response research in the future.

Fourth, as alluded to several times throughout this discussion, an important next step in advancing person-fit research would be identifying method(s) that provide optimal cutoffs. Many of the results from this dissertation are likely due to the use of a conservative and conventional cutoff. These person-fit statistics could show better identification rates if the alpha levels were modified to allow for more Type I errors. Notably, it would be important to find a cutoff that provided the biggest benefit to sensitivity while still maintaining acceptable specificity levels. A study investigating the optimal cutoff for these nonparametric person-fit statistics would need to simulate data, find the classification results using the person-fit statistics, and identify a cutoff method that provides an acceptable level of Type I and Type II errors. The results could be found with a ROC analysis or other methods commonly used to examine classification accuracy.

This dissertation provided evidence that dimensionality has an impact on the accuracy of nonparametric person-fit statistics in polytomous response data. However, simulated response data were based on an underlying measurement model with correlated factors. It is unknown whether orthogonal factors or the magnitude of the factor intercorrelations from the underlying measurement model would impact the accuracy of these person-fit statistics. While unrealistic, orthogonal factors are often used in low-stakes surveys in both education and psychology research. It is possible that a measurement model with orthogonal factors would decrease the accuracy of these person-fit statistics. Additionally, it is possible the person-fit statistics would become more accurate as the intercorrelations between factors increased in magnitude. A study of

this nature would be of particular interest when focused on H_i^T, as the accuracy of H_i^T was most impacted by dimensionality.

Finally, the robustness of these person-fit statistics to the violation of the unidimensionality assumption needs additional investigation. This study showed that the violation of unidimensionality does impact the accuracy of these person-fit statistics, but not in every condition. The nature of this robustness and the impact of dimensionality is still unclear. A study investigating the robustness of the person-fit statistics to dimensionality would need to calculate the bias (i.e., systematic error) introduced into the person-fit estimates from differing levels of dimensionality. Doing so would require a simulation where data were generated using specified parameters. For example, a data set where H_i^T is a known value. Keeping all else equal, dimensionality could be introduced as a factor, H_i^T estimated, and differences catalogued. Over the course of many replications, a pattern would start to emerge giving information about how much bias is introduced at various dimensionalities.

REFERENCES

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Waveland Press.

Armstrong, R. D., Stoumbos, Z. G., Kung, M. T., & Shi, M. (2007). On the performance of the l_z person-fit statistic. *Practical Assessment, Research and Evaluation, 12*. Retreived from https://scholarworks.umass.edu/pare/vol12/iss1/16.

Baumgartner, H. & Steenkamp, J. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*, 143-156.

Beck, M. F. (2015). *The Multidimensional Social Anxiety Response Inventory (MSARI): Development and preliminary examination of psychometric properties* (Publication No. 1588285) [Master's thesis, The University of Texas at San Antonio]. ProQuest Dissertations and Theses.

Beck, M. F., Albano, A. D., & Smith, W. M. (2019). Person-fit as an index of inattentive responding: A comparison of methods using polytomous survey data. *Applied Psychological Measurement, 43*, 374-387.

Becker, W. M. (1976). Biasing effect of respondents' identification on responses to social desirability scale: A warning to researchers. *Psychological Reports, 39*, 756-758.

Belli, R. F., Lee, E. H., Stafford, F. P., & Chou, C. (2004). Calendar and question-list survey methods: Association between interviewer behaviors and data quality. *Journal of Official Statistics, 20*, 185-218.

Bilder, C. R. & Loughin, T. M. (2015). *Analysis of Categorical Data with R*. CRC Press.

Blau, G. & Katerberg, R. (1982). Agreeing response set: Statistical nuisance of meaningful personality concept. *Perceptual and Motor Skills, 54*, 851-857.

Böckenholt, U. (2012). Modeling multiple response processes in judgement and choice. *Psychological Methods, 17,* 665-678.

Böckenholt, U. (2017). Measuring response styles in Likert items. *Psychological Methods, 22,* 69-83.

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

Bolt, D. M. & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement, 33,* 335-352.

Bornstein, R. F. (1990). Publication politics, experimenter bias and the replication process in social science research. *Journal of Social Behavior and Personality, 5*(4), 71-81.

Buderer, N. M. (1996). Statistical methodology: I. Incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity. *Academic Emergency Medicine, 3*(9), 895-900.

Bulut, O., & Sünbül, Ö. (2017). Monte Carlo simulation studies in item response theory with the R programming language. *Journal of Measurement and Evaluation in Education and Psychology, 8,* 266-287.

Burton, S. & Blair, E. (1991). Task conditions, response formulation process, and response accuracy for behavioral frequency questions in surveys. *Public Opinion Quarterly, 55,* 50-79.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48,* 1-29.

Choi, B. C. K. & Pak, A. W. P. (2005). A Catalog of biases in questionnaires. *Preventing Chronic Disease, 2,* 1-13.

Clark, M. E., Gironda, R. J., & Young, R. W. (2003). Detection of back random responding: Effectiveness of MMPI-2 and personality assessment inventory validity indices. *Psychological Assessment, 15*, 223-234.

Cohen, J. (1992a). A power primer. *Psychological Bulletin, 112*(1), 155-159.

Cohen, J. (1992b). Statistical power analysis. *Current Directions in Psychological Science, 1*(3), 98-101.

Couch, A., & Keniston, K. (1960). Yea-sayers and nay-sayers: Agreeing response set as a personality variable. *Journal of Abnormal and Social Psychology, 60,* 151-174.

Couper, M. P., Traugott, M. W., & Lamias, M. J. (2001). Web survey design and administration. *Public Opinion Quarterly, 65*, 230-253.

Cox, E. P., III (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research, 17*(4), 407-422.

Credé, M. (2010). Random Responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement, 70*, 596-612.

Crişan, D. R., Tendeiro, J. N., & Meijer, R. R. (2017). Investigating the practical consequences of model misfit in unidimensional IRT models. *Applied Psychological Measurement, 41*, 439-455.

Crown, D. P. & Marlowe, D. (1960). A new scale of social desirability. *Journal of Consulting Psychology, 24*, 349-354.

De Ayala, R. J. (1994). The influence of multidimensionality on the graded response model. *Applied Psychological Measurement, 18*, 155-170.

De Ayala, R. J. (2009). *Theory and practice of item response theory*. The Guilford Press.

DeCastellarnau, A. (2018). A classification of response scale characteristics that affect data quality: A literature review. *Quality and Quantity, 52*, 1523-1559.

De Jong, M. G., Steenkamp, J. E. M., Fox, J., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research, 45*, 105-115.

DeVellis, R. F. (2012) *Scale development: Theory and applications*. Sage Publications, Inc.

Dimitrov, D. M. & Smith, R. M. (2006). Adjusted Rasch person-fit statistics. *Journal of Applied Measurment, 7*, 170-183.

Doty, D. H. & Glick, W. H. (1998). Common methods bias: Does common methods variance really bias results? *Organizational Research methods, 1*, 374-406.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67-86.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional Item Response Theory models to multidimensional data. *Applied Psychological Measurement, 7*(2), 189-199.

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research*. Dryden Press.

Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous items scores. *Applied Psychological Measurement, 32*, 224-247.

Emons, W. H. M., Miejer, R. R., & Sijtsma, K. (2002). Comparing simulated and theoretical sampling distributions of the U3 person-fit statistics. *Applied Psychological Measurement, 26*, 88-108.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2005). Global, local, and graphical person-fit analysis using person-response functions. *Psychological Methods, 10*, 101-119.

Finch, H. (2011). Multidimensional item response theory parameter estimation with nonsimple structure items. *Applied Psychological Measurement, 35*(1), 67-82.

Fischer, D. G & Fick, C. (1993). Measuring social desirability: Short forms of the Marlowe-Crowne social desirability scale. *Educational and Psychological Measurement, 53*, 417-424.

Fowler, F. J., Jr. (2009). *Survey research methods*. Sage Publications, Inc.

Frias-Navarro, D., Pascual-Llobell, J., Pascual-Soler, M., Perezgonzalez, J., & Berrios-Riquelme, J. (2020). Replication crisis or an opportunity to improve scientific production? *European Journal of Education, 55*, 618-631.

Greenleaf, E. A. (1992). Measuring extreme response style. *The Public Opinion Quarterly, 56*, 328-351.

Gold, B. (1975). "A comment on Shulman's 'Comparison of two scales on extremity response bias'". *Public Opinion Quarterly, 39*, 123-124.

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society, 46*(2), 149-192.

Groves, R. M. (1987). Research on survey data quality. *The Public Opinion Quarterly, 51*, S156-S172.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed. pp. 147-200). Macmillan.

Hamilton, D. L. (1968). Personality attributes associated with extreme response style. *Psychological Bulletin, 69*, 192-203.

Harnisch, D. L. & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18*, 133-146.

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics, 11*(2), 91-115.

Hauser, D. J. & Schwarz, N. (2016). Attentive turkers: MTurk participants perform better attention checks than do subject pool participants. *Behavioral Research Methods, 48*, 400-407.

He, J., Bartram, D., Inceoglu, I., & van de Bijver, F. J. R. (2014). Response styles and personality traits: A multilevel analysis. *Journal of Cross-Cultural Psychology, 45*, 1028-1045.

Hernández, A., Drasgow, F., & González-Romá, V. (2004). Investigating the functioning of a middle category by means of a mixed-measurement model. *Journal of Applied Psychology, 89*, 687-699.

Hinz, A., Michalski, D., Schwarz, R., & Herzberg, P. Y. (2007). The acquiescence effect in responding to a questionnaire. *GMS Psycho-Social-Medicine, 4*, 1-9.

Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. *Personality and Social Psychology Bulletin, 30*, 161-172

Hong, M. Steedle, J. T., & Cheng, Y. (2020). Methods of detecting insufficient effort responding: Comparisons and practical recommendations. *Educational and Psychological Measurement, 80*, 312-345.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *The Journal of Business and Psychology, 27*, 99-114.

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology, 100*, 828-845.

Jaccard, J. & Wan, C. K. (1995). Measurement error in the analysis of interaction effects between continuous predictors using multiple regression: Multiple indicator and structural equation approaches. *Quantitative Methods in Psychology*, *117*(2), 348-357.

Jackson, D. N. (1976). *The appraisal of personal reliability* [Paper presentation]. The Society of Multivariate Experimental Psychology, University Park, PA, United States.

Jackson, D. N. & Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin, 55*, 243-252.

Jackson, D. N. & Messick, S. (1962). Response styles on the MMPI: Comparison of clinical and normal samples. *Journal of Abnormal and Social Psychology, 65*, 285-299.

Jacobson, L. I., Kellogg, R. W., Cauce, A. M., & Slavin, R. S. (1977). A multidimensional social desirability inventory. *Bulletin of the Psychonomic Society, 9*, 109-110.

Jeon, M. & De Boeck, P. (2019). Evaluation on types of invariance in studying extreme response bias with an IRTree approach. *British Journal of Mathematical and Statistical Psychology, 72*, 517-537.

Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in Psychology, 7*, 1-10.

Jin K., Chen, H., & Wang, W. (2018). Mixture item response models for inattentive responding behavior. *Organizational Research Methods, 21*, 197-225.

Jin, K., & Wang, W. (2014). Generalized IRT models for extreme response style. *Educational and Psychological Measurement, 74*, 116-138.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality, 39*, 103-129.

Jones, E. E., & Sigall, H. (1971) The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin, 76*, 349-364.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*, 277-298.

Keren, G., & Lewis, C. (1979). Partial omega squared for ANOVA designs. *Educational and Psychological Measurement, 39*, 119-128.

Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of behavioral research* (4th ed.). Cengage Learning.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8*, 147-154.

Knowles, E. S., & Condon, C. A. (1999). Why people say "yes": A dual-process theory of acquiescence. *Journal of Personality and Social Psychology, 77*, 379-386.

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement, 67*, 606-619.

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research* (2nd ed., pp. 263-313). Emerald Group Publishing Limited.

Kwak, D., Holtkamp, P., & Kim, S. S. (2019). Measuring and controlling social desirability bias: Applications in information systems research. *Journal of the Association for Information Systems, 20*, 317-345.

Levine, M. V. & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*, 269-290.

Lietz, P. (2010) Research into questionnaire design: A summary of the literature. *International Journal of Market Research, 52*, 249-272.

Ligtvoet, R. van der Ark, L. A., te Marvelde, J. M., & Sijtsma, K. (2010). Investigating an invariant ordering for polytomously scored items. *Educational and Psychological Measurement, 70*, 578-595.

Loevinger, J. (1948). The technic of homogenous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin, 45*, 507-530.

Magis, D., Raîche, G., & Béland, S. (2012). A didactic presentation of Snijder's l_z* index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics, 37*, 57-81.

Maniaci, M. R. & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality, 48*, 61-83.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist, 70*(6), 487-498.

McDonald, R. P. (1999). *Test theory: A unified treatment*. L. Erlbaum Associates.

Mckibben, W. B. & Silvia, P. J. (2015). Evaluating the distorting effects of inattentive responding and social desirability on self-report scales in creativity and the arts. *The Journal of Creative Behavior, 51*, 57-69.

Meade, A. W. & Craig, B. (2012) Identifying careless responses in survey data. *Psychological Methods, 17*, 437-455.

Meade, A.W. & Pappalardo, G. (2013). *Predicting careless responses and attrition in survey data with personality* [Paper presentation]. The 28th Annual Meeting of the Society for Industrial and Organizational Psychology, Houston, TX, United States.

Meehl, P. E. & Hathaway, S. R. (1946). The K factor as a suppressor variable in the Minnesota Multiphasic Personality Inventory. *Journal of Applied Psychology, 30,* 525-564.

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18,* 311-314.

Meijer, R. R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods, 8*(1), 72-87.

Meijer, R. R. & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods, 9,* 354-368.

Meijer, R. R. & Sijtsma, K. (1995). Detection of aberrant item score patterns. *Applied Measurement in Education, 8,* 261-272.

Meijer, R. R. & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25,* 107-135.

Meijer, R. R. & Tenderio, J. N. (2012). The use of the l_z and l_z* person-fit statistics and problems derived from model misspecification. *Journal of Educational and Behavioral Statistics, 37,* 758-766.

Meisenberg, G. & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences, 44,* 1539-1550.

Messick, S. (1991). Psychology and methodology of response styles. *Improving Inquiry in Social Science: A Volume in Honor of Lee J. Cronbach,* 161-200.

Meyers, J. L., Miller, G. E., & Way, W. D. (2008). Item position and item difficulty changes in an IRT based common item equating design. *Applied Measurement in Education, 22*, 38-60.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Mouton & Co.

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351-365). Springer.

Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 369-380). Springer.

Moors, G. (2008). Exploring the effect of a middle response category on response style in attitude measurement. *Quality & Quantity, 42*, 779-794.

Morawski, J. (2019). The replication crisis: How might philosophy and theory of psychology be of use? *Theoretical and Philosophical Psychology, 39*(4), 218-238.

Mousavi, A., Cui, Y., & Rogers, T. (2019). An examination of different methods of setting cutoff values in person fit research. *International Journal of Testing, 19*, 1-22.

Nichols, D. S., Greene, R. L., & Schmolk, P. (1989). Criteria for assessing inconsistent patterns of item endorsement on the MMPI: Rationale, development, and empirical trials. *Journal of Clinical Psychology, 45*, 239-250.

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality, 63*, 1-11.

Paulus, D. L. (1984). Two-component models of socially desirable responding. *Personality Processes and Individual Differences, 46*, 598-609.

Paulus, D. L. (1988). Balanced Inventory of Desirable Responding (BIDR). *Acceptance and Commitment Therapy. Measures Package, 41*, 41-43.

Paulus, D. L. (1991). Measurement and control of response bias. In J.P. Robinson, P.R. Shaver, & L.S. Wrightsman (Eds.), *Measures of social psychological attitudes, Vol. 1*. Academic Press.

Proctor, C. H. (1970). A probabilistic formulation and statistical analysis of Guttman scaling. *Psychometrika, 35*, 73-78.

R Core Team (2021). *R: A language and environment for statistical computing* [Computer Software]. R Foundation for Statistical Computing. https://www.R-project.org/.

Revilla, M. A., Saris, W. E., & Krosnick, J. A. (2014). Choosing the number of categories in agree-disagree scales. *Sociological Methods and Research, 43*, 73-97.

Rossi, P. E., Gilula, Z., & Allenby, G. M. (2001). Overcoming scale usage heterogeneity: A Bayesian hierarchical approach. *Journal of the American Statistical Association, 96*, 20-31.

Sauro, J. & Lewis, J. R. (2011, May 7-12). *When designing usability questionnaires, does it hurt to be positive?* [Conference session]. ACM CHI Conference on Human Factors in Computing Systems, Vancouver, BC, Canada. https://dl.acm.org/doi/pdf/10.1145/1978942.1979266?casa_token=eYAKkj829Sc

AAAAA:YibDhNsQ35oGnD5rreNcdlOpHe-kH3D9fA9gAl6lrcaU0zUhnZyNr2bwjoiF-UohAtxD1rdwJgU

Schriesheim, C. A., & Hill, K. D. (1981). Controlling acquiescence response bias by item reversals: The effect on questionnaire validity. *Educational and Psychological Measurement, 41,* 1101-1114.

Schurr, K. T. & Henriksen, L. W. (1983). Effects of item sequencing and grouping in low-inference type questionnaires. *Journal of Educational Measurement, 20,* 379-391.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54,* 93-105.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton Mifflin Company.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. McGraw-Hill Book Company, Inc.

Sijstma, K. & Meijer, R. R. (1992). A method for investigating the intersections of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement, 16,* 149-157.

Sijstma, K. & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Sage Publications, Inc.

Singleton, R.A., & Straits, B.C. (2009). *Approaches to social research.* Oxford University Press.

Sinharay, S. (2016). Asymptotic corrections of the standardized extended caution indices. *Applied Psychological Measurement, 40*, 418-433.

Sinharay, S. (2017). Are the nonparametric person-fit statistics more powerful than their parametric counterparts? Revisiting the simulation in Karabatsos (2003). *Applied Measurement in Education, 30*, 314-328.

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika, 66*, 331-342.

Soland, J., Wise, S. L., & Gao, L. (2019). Identifying disengaged survey responses: New evidence using response time metadata. *Applied Measurement in Education, 32*, 151-165.

St-Onge, C., Valois, P., Abdous, B., & Germain, S. (2011). Accuracy of person-fit statistics: A monte carlo study of the influence of aberrance rates. *Applied Psychological Measurement, 35*, 1-14.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology.* Jossey-Bass.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49*, 95-110.

Tatsuoka, K. K. & Linn, R. L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement, 7*, 81-96.

Tendeiro, J. N. & Meijer, R. R. (2014). Detection of invalid test scores: The usefulness of simple nonparametric statistics. *Journal of Educational Measurement, 51*, 239-259.

Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. (2016). PerFit: An R Package for Person-Fit Analysis in IRT. *Journal of Statistical Software, 74*(5), 1-27.

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements*. The Science Press.

van de Vijver, F. J. & He, J. (2014). *Report on social desirability, midpoint and extreme responding in TALIS 2013* (Report No. 107). OECD publishing.

van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*, 1-19.

van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software, 48*, 1-27.

van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology, 13*, 267-298.

van Krimpen-Stoop, E. M. L. A. & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement, 23*, 327-345.

van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. *PLOS ONE, 8*, 1-7.

Velez, P. & Ashworth, S. D. (2007). The impact of item readability on the endorsement of the midpoint response in surveys. *Survey Research Methods, 1*, 69-74.

Wang, W., Wilson, M., & Shih, C. (2006). Modeling randomness in judging rating scales with a random-effects rating scale model. *Journal of Educational Measurement, 43*, 335-353.

Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative model. *Psychological methods, 18,* 320-334.

Weijters, B., Cabooter, E., Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing, 27,* 236-247.

Wetzel, E., Böhnke, J. R., & Brown, A. (2016). Response biases. In F.T. Leong, D. Bartram, F. Cheung, K.F. Geisinger, & D. Illiescu (Eds.), *The ITC international handbook of testing and assessment* (pp. 349-363). Oxford University Press.

Wetzel, E., Carstensen, C. H., & Böhnke, J. R. (2013). Consistency of extreme response styles and non-extreme response style across traits. *Journal of Research in Personality, 47,* 178-189.

Wiggins, J. S. (1964). Convergence among stylistic response measures from objective personality tests. *Educational and Psychological Measurement, 24,* 551-562.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18,* 163-183.

Zhang, Y. & Wang, Y. (2020). Validity of three IRT models for measuring and controlling extreme and midpoint response styles. *Frontiers in Psychology, 11,* 1-10.

APPENDIX A

Table A1

Covariance Matrix and Means for Data Simulation

	Mean	**D1**	**D2**	**D3**	**D4**
D1	*0.03*	1.02			
D2	*0*	0.49	1.11		
D3	*0.09*	0.31	0.38	0.95	
D4	*0.02*	0.43	0.89	0.40	0.96

Table A2

Uniform Distributions of Item Boundary Parameters by Condition

k	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5
4	*U*(-13.71, -5.19)	*U*(-5.19, -2.51)	*U*(-2.51, 2.42)		
5	*U*(-13.71, -5.81)	*U*(-5.81, -2.19)	*U*(-2.19, 2.42)	*U*(1.5, 3)	
6	*U*(-13.71, -6.68)	*U*(-6.68, -4.78)	*U*(-4.78, -3.09)	*U*(-3.09, -1.42)	*U*(-1.42, 2.42)

Note. Cat stands for Category

Table A3

Item Discrimination Parameters that will be used for Data Generation

a1	a2	a3	a4
2.27	0.16	0.15	0.28
2.42	0.31	0.34	0.10
2.70	0.27	0.22	0.19
3.17	0.15	0.20	0.18
2.18	0.38	0.28	0.34
3.15	0.38	0.28	0.18
3.22	0.14	0.14	0.32
2.82	0.35	0.19	0.37
2.78	0.24	0.27	0.38
1.98	0.26	0.29	0.12
2.18	0.27	0.25	0.33
2.14	0.17	0.25	0.19
2.86	0.33	0.26	0.13
2.43	0.15	0.27	0.39
2.97	0.22	0.36	0.22
2.59	0.36	0.35	0.24
2.90	0.39	0.13	0.39
3.28	0.17	0.31	0.28
2.43	0.23	0.37	0.39
2.98	0.12	0.18	0.33
3.20	0.30	0.17	0.31
2.19	0.22	0.10	0.40
2.81	0.35	0.14	0.25
2.07	0.15	0.13	0.25
2.27	0.20	0.17	0.29
2.44	0.25	0.34	0.35
1.91	0.14	0.28	0.24
2.43	0.21	0.37	0.35
3.11	0.39	0.27	0.25
2.37	0.14	0.33	0.26
2.57	0.10	0.21	0.27
2.73	0.15	0.21	0.17
2.59	0.34	0.15	0.36
2.16	0.36	0.24	0.30
3.05	0.25	0.18	0.24
2.83	0.29	0.20	0.39
0.34	**4.36**	0.37	0.24
0.13	**3.23**	0.16	0.29

0.32	**4.00**	0.27	0.22
0.22	**2.96**	0.16	0.10
0.35	**4.63**	0.18	0.38
0.29	**3.25**	0.34	0.17
0.33	**2.89**	0.15	0.27
0.27	**2.99**	0.27	0.15
0.26	**4.56**	0.23	0.37
0.34	**4.26**	0.18	0.13
0.11	**4.62**	0.11	0.37
0.24	**3.36**	0.13	0.37
0.32	**3.67**	0.19	0.32
0.31	**4.29**	0.34	0.27
0.24	**2.67**	0.17	0.22
0.36	**2.69**	0.16	0.32
0.23	**4.03**	0.36	0.37
0.17	**4.53**	0.40	0.34
0.12	**3.21**	0.35	0.35
0.13	**4.29**	0.37	0.23
0.19	**4.24**	0.24	0.15
0.26	**4.65**	0.17	0.15
0.30	**3.89**	0.14	0.37
0.22	**4.09**	0.18	0.32
0.37	**4.21**	0.34	0.27
0.19	**4.44**	0.12	0.12
0.24	**3.91**	0.34	0.36
0.20	**3.18**	0.13	0.37
0.30	**4.38**	0.33	0.17
0.18	**3.54**	0.19	0.29
0.24	**3.44**	0.33	0.12
0.33	**3.59**	0.26	0.25
0.13	**3.10**	0.21	0.34
0.36	**2.79**	0.13	0.39
0.20	**3.21**	0.33	0.20
0.35	**3.28**	0.33	0.29
0.20	0.11	**6.28**	0.22
0.20	0.16	**6.52**	0.20
0.24	0.16	**4.81**	0.35
0.37	0.33	**4.94**	0.31
0.36	0.19	**3.47**	0.20
0.22	0.36	**3.48**	0.23
0.33	0.22	**5.84**	0.18
0.39	0.27	**5.71**	0.14
0.23	0.21	**5.83**	0.37

0.31	0.30	**5.34**	0.31
0.22	0.11	**4.29**	0.27
0.20	0.22	**2.88**	0.38
0.33	0.16	**6.48**	0.37
0.16	0.36	**6.04**	0.12
0.31	0.39	**3.66**	0.11
0.14	0.20	**4.73**	0.40
0.17	0.32	**5.27**	0.16
0.14	0.20	**6.35**	0.38
0.17	0.39	**2.90**	0.16
0.12	0.22	**3.87**	0.14
0.29	0.21	**4.51**	0.26
0.36	0.27	**6.00**	0.18
0.33	0.24	**6.16**	0.21
0.34	0.16	**3.81**	0.13
0.24	0.23	**4.08**	0.33
0.22	0.13	**4.01**	0.23
0.34	0.13	**3.55**	0.11
0.28	0.23	**5.43**	0.31
0.30	0.16	**5.76**	0.18
0.21	0.23	**5.44**	0.29
0.18	0.39	**3.65**	0.18
0.40	0.35	**5.56**	0.26
0.29	0.19	**5.15**	0.24
0.16	0.28	**4.15**	0.27
0.14	0.37	**3.01**	0.30
0.24	0.24	**4.38**	0.16
0.38	0.14	0.12	**3.79**
0.28	0.14	0.19	**3.45**
0.39	0.11	0.20	**2.36**
0.32	0.32	0.12	**2.61**
0.21	0.21	0.15	**2.39**
0.23	0.27	0.15	**2.81**
0.14	0.35	0.37	**3.34**
0.10	0.34	0.31	**2.84**
0.31	0.36	0.35	**2.69**
0.13	0.13	0.30	**3.41**
0.23	0.39	0.16	**3.66**
0.29	0.27	0.24	**1.95**
0.40	0.11	0.32	**2.09**
0.25	0.17	0.35	**2.60**
0.25	0.39	0.35	**3.54**
0.15	0.37	0.16	**2.34**

0.33	0.17	0.28	**2.00**
0.24	0.33	0.23	**3.28**
0.25	0.27	0.29	**3.80**
0.16	0.19	0.23	**2.21**
0.17	0.31	0.36	**2.93**
0.28	0.20	0.20	**2.51**
0.27	0.16	0.19	**2.63**
0.12	0.38	0.25	**2.19**
0.11	0.11	0.20	**2.51**
0.29	0.39	0.17	**2.18**
0.38	0.19	0.35	**3.46**
0.28	0.30	0.26	**3.60**
0.27	0.26	0.13	**3.23**
0.26	0.35	0.13	**2.37**
0.40	0.16	0.27	**2.68**
0.25	0.22	0.26	**1.83**
0.30	0.15	0.25	**2.95**
0.28	0.19	0.20	**3.17**
0.17	0.29	0.30	**2.44**
0.18	0.19	0.36	**3.73**

Note. Primary loadings are bolded.

Table A4

Specifications of Static Data Sets

Length	Number of Dimensions	Items from Factor 1	Items from Factor 2	Items from Factor 3	Items from Factor 4
12	1	12	0	0	0
	2	6	6	0	0
	3	4	4	4	0
	4	3	3	3	3
36	1	36	0	0	0
	2	18	18	0	0
	3	12	12	12	0
	4	9	9	9	9

*_Note_. Dim = dimension

Table A5

Possible Results of Classification

		True Status	
		Aberrant	Not Aberrant
Classification Result	Aberrant	*True Positive* (1- β)	*False Positive* (α)
	Not Aberrant	*False Negative* (β)	*True Negative* (1- α)

Table A6

Classification Accuracy of Guttman Errors across Simulation Conditions for Acquiescence Responding

Dimension	**NPV**	**PPV**	**Sensitivity**	**Specificity**
1	0.90	0.07	2.74	94.58
2	0.90	0.12	6.36	94.99
3	0.90	0.13	5.80	94.91
4	0.90	0.11	5.36	94.98
Response Categories	**NPV**	**PPV**	**Sensitivity**	**Specificity**
4	0.90	0.06	3.22	94.44
5	0.90	0.18	9.24	95.37
6	0.90	0.08	4.12	94.73
7	0.90	0.10	5.36	94.92
Test Length	**NPV**	**PPV**	**Sensitivity**	**Specificity**
12	0.90	0.07	3.77	94.60
36	0.90	0.14	7.20	95.13

Table A7

Classification Accuracy of H^T_i across Simulation Conditions for Acquiescence Responding

Dimension	**NPV**	**PPV**	**Sensitivity**	**Specificity**
1	0.90	0.18	9.24	95.38
2	0.90	0.15	7.44	95.18
3	0.90	0.13	6.58	95.09
4	0.90	0.19	9.38	95.40
Response Categories	**NPV**	**PPV**	**Sensitivity**	**Specificity**
4	0.90	0.09	4.41	94.85
5	0.90	0.09	4.78	94.89
6	0.90	0.16	8.20	95.28
7	0.91	0.30	15.24	96.04
Test Length	**NPV**	**PPV**	**Sensitivity**	**Specificity**
12	0.90	0.15	7.42	95.19
36	0.90	0.18	8.89	95.34

Table A8

Classification Accuracy of U3 across Simulation Conditions for Acquiescence Responding

Dimension	**NPV**	**PPV**	**Sensitivity**	**Specificity**
1	0.90	0.08	5.01	94.71
2	0.90	0.10	5.71	94.87
3	0.90	0.10	5.21	94.81
4	0.90	0.08	4.53	94.81
Response Categories	**NPV**	**PPV**	**Sensitivity**	**Specificity**
4	0.90	0.07	4.22	94.68
5	0.90	0.14	8.34	94.94
6	0.90	0.09	4.59	94.84
7	0.90	0.06	3.32	94.74
Test Length	**NPV**	**PPV**	**Sensitivity**	**Specificity**
12	0.90	0.06	3.59	94.60
36	0.90	0.12	6.65	95.00

Table A9

Classification Accuracy of Guttman Errors across Simulation Conditions for Disacquiescence Responding

Dimension	**NPV**	**PPV**	**Sensitivity**	**Specificity**
1	0.91	0.35	17.90	96.24
2	0.91	0.23	12.17	95.66
3	0.90	0.17	8.60	95.34
4	0.91	0.22	11.14	95.57
Response Categories	**NPV**	**PPV**	**Sensitivity**	**Specificity**
4	0.91	0.34	17.61	96.27
5	0.91	0.31	15.73	96.10
6	0.90	0.18	9.58	95.30
7	0.90	0.13	6.88	95.13
Test Length	**NPV**	**PPV**	**Sensitivity**	**Specificity**
12	0.91	0.23	12.05	95.60
36	0.91	0.25	12.85	95.80

Table A10

Classification Accuracy of H^T_i across Simulation Conditions for Disacquiescence Responding

Dimension	**NPV**	**PPV**	**Sensitivity**	**Specificity**
1	0.90	0.09	4.64	94.88
2	0.90	0.07	3.73	94.79
3	0.90	0.07	3.77	94.75
4	0.90	0.09	4.53	94.86
Response Categories	**NPV**	**PPV**	**Sensitivity**	**Specificity**
4	0.90	0.14	7.03	95.13
5	0.90	0.11	5.37	94.99
6	0.90	0.05	2.54	94.65
7	0.90	0.03	1.73	94.52
Test Length	**NPV**	**PPV**	**Sensitivity**	**Specificity**
12	0.90	0.06	3.12	94.70
36	0.90	0.10	5.21	94.94

Table A11

Classification Accuracy of U3 across Simulation Conditions for Disacquiescence Responding

Dimension	NPV	PPV	Sensitivity	Specificity
1	0.91	0.21	11.01	95.59
2	0.91	0.23	11.71	95.69
3	0.90	0.18	9.44	95.45
4	0.91	0.25	12.66	95.78
Response Categories	**NPV**	**PPV**	**Sensitivity**	**Specificity**
4	0.91	0.28	14.47	95.98
5	0.91	0.24	12.52	95.77
6	0.90	0.18	9.30	95.41
7	0.90	0.17	8.53	95.34
Test Length	**NPV**	**PPV**	**Sensitivity**	**Specificity**
12	0.91	0.22	11.44	95.65
36	0.91	0.21	10.97	95.60

Table A12

Classification Accuracy of Guttman Errors across Simulation Conditions for Midpoint Responding

Dimension	NPV	PPV	Sensitivity	Specificity
1	0.90	0.00	0.08	94.38
2	0.90	0.00	0.16	94.41
3	0.90	0.00	0.16	94.38
4	0.90	0.00	0.07	94.39
Response Categories	**NPV**	**PPV**	**Sensitivity**	**Specificity**
4	0.90	0.00	0.07	94.39
5	0.90	0.00	0.07	94.36
6	0.90	0.00	0.02	94.37
7	0.90	0.01	0.31	94.43
Test Length	**NPV**	**PPV**	**Sensitivity**	**Specificity**
12	0.90	0.00	0.14	94.38
36	0.90	0.00	0.09	94.40

Table A13

Classification Accuracy of H^T_i across Simulation Conditions for Midpoint Responding

Dimension	**NPV**	**PPV**	**Sensitivity**	**Specificity**
1	0.92	0.39	20.10	96.59
2	0.91	0.20	10.35	95.50
3	0.90	0.17	8.90	95.33
4	0.91	0.22	11.13	95.62
Response Categories	**NPV**	**PPV**	**Sensitivity**	**Specificity**
4	0.91	0.20	10.39	95.51
5	0.91	0.21	10.81	95.58
6	0.91	0.31	16.03	96.14
7	0.91	0.26	13.25	95.82
Test Length	**NPV**	**PPV**	**Sensitivity**	**Specificity**
12	0.91	0.20	10.29	95.50
36	0.91	0.29	14.95	96.02

Table A14

Classification Accuracy of U3 across Simulation Conditions for Midpoint Responding

Dimension	**NPV**	**PPV**	**Sensitivity**	**Specificity**
1	0.90	0.01	0.39	94.39
2	0.90	0.00	0.11	94.44
3	0.90	0.00	0.23	94.41
4	0.90	0.00	0.14	94.41
Response Categories	**NPV**	**PPV**	**Sensitivity**	**Specificity**
4	0.90	0.00	0.21	94.44
5	0.90	0.00	0.12	94.41
6	0.90	0.00	0.07	94.37
7	0.90	0.01	0.47	94.42
Test Length	**NPV**	**PPV**	**Sensitivity**	**Specificity**
12	0.90	0.00	0.22	94.41
36	0.90	0.00	0.21	94.41

Table A15

Classification Accuracy of Guttman Errors across Simulation Conditions for Extreme Responding

Dimension	NPV	PPV	Sensitivity	Specificity
1	0.92	0.55	27.93	97.43
2	0.92	0.48	24.74	97.11
3	0.92	0.50	25.95	97.19
4	0.92	0.53	27.07	97.35
Response Categories	**NPV**	**PPV**	**Sensitivity**	**Specificity**
4	0.93	0.57	29.34	97.59
5	0.93	0.63	31.89	97.92
6	0.92	0.55	28.07	97.44
7	0.91	0.32	16.40	96.15
Test Length	**NPV**	**PPV**	**Sensitivity**	**Specificity**
12	0.92	0.42	21.81	96.74
36	0.93	0.61	31.04	97.80

Table A16

Classification Accuracy of H^T_i across Simulation Conditions for Extreme Responding

Dimension	**NPV**	**PPV**	**Sensitivity**	**Specificity**
1	0.90	0.12	5.92	95.01
2	0.90	0.08	3.95	94.78
3	0.90	0.09	4.63	94.93
4	0.90	0.10	4.91	94.93
Response Categories	**NPV**	**PPV**	**Sensitivity**	**Specificity**
4	0.90	0.13	6.53	95.11
5	0.90	0.10	5.34	94.96
6	0.90	0.09	4.40	94.84
7	0.90	0.06	3.15	94.75
Test Length	**NPV**	**PPV**	**Sensitivity**	**Specificity**
12	0.90	0.08	4.06	94.83
36	0.90	0.11	5.64	94.99

Table A17

Classification Accuracy of U3 across Simulation Conditions for Extreme Responding

Dimension	NPV	PPV	Sensitivity	Specificity
1	0.92	0.48	24.95	97.06
2	0.92	0.45	23.10	96.88
3	0.92	0.46	23.68	96.94
4	0.92	0.46	23.73	96.97
Response Categories	**NPV**	**PPV**	**Sensitivity**	**Specificity**
4	0.92	0.50	25.72	97.15
5	0.93	0.63	32.50	97.90
6	0.92	0.49	25.06	97.10
7	0.91	0.24	12.18	95.70
Test Length	**NPV**	**PPV**	**Sensitivity**	**Specificity**
12	0.91	0.36	18.50	96.36
36	0.93	0.57	29.23	97.56

Table A18

Classification Accuracy of Guttman Errors across Simulation Conditions for Social Desirability Responding

Dimensions	**NPV**	**PPV**	**Sensitivity**	**Specificity**
1	0.90	0.11	5.78	94.92
2	0.90	0.14	7.25	95.19
3	0.90	0.15	7.79	95.25
4	0.90	0.15	7.79	95.23
Response Categories	**NPV**	**PPV**	**Sensitivity**	**Specificity**
4	0.90	0.09	4.79	94.91
5	0.90	0.15	7.68	95.23
6	0.90	0.12	6.19	94.99
7	0.91	0.19	9.94	95.47
Test Length	**NPV**	**PPV**	**Sensitivity**	**Specificity**
12	0.90	0.12	6.01	95.01
36	0.90	0.16	8.29	95.29

Table A19

Classification Accuracy of H^T_i across Simulation Conditions for Social Desirability Responding

Dimension	**NPV**	**PPV**	**Sensitivity**	**Specificity**
1	0.90	0.17	8.46	95.32
2	0.90	0.17	8.43	95.30
3	0.90	0.19	9.40	95.42
4	0.91	0.22	11.32	95.61
Response Categories	**NPV**	**PPV**	**Sensitivity**	**Specificity**
4	0.90	0.09	4.65	94.90
5	0.90	0.10	4.95	94.93
6	0.91	0.22	11.04	95.59
7	0.91	0.33	16.97	96.24
Test Length	**NPV**	**PPV**	**Sensitivity**	**Specificity**
12	0.90	0.14	7.18	95.18
36	0.91	0.23	11.62	95.64

Table A20

Classification Accuracy of U3 across Simulation Conditions for Social Desirability Responding

Dimension	**NPV**	**PPV**	**Sensitivity**	**Specificity**
1	0.90	0.10	5.56	94.90
2	0.90	0.11	5.64	94.96
3	0.90	0.14	7.22	95.01
4	0.90	0.13	6.58	95.08
Response Categories	**NPV**	**PPV**	**Sensitivity**	**Specificity**
4	0.90	0.08	4.20	94.59
5	0.90	0.11	6.02	94.97
6	0.90	0.11	5.52	94.98
7	0.90	0.18	9.27	95.41
Test Length	**NPV**	**PPV**	**Sensitivity**	**Specificity**
12	0.90	0.10	5.48	94.86
36	0.90	0.14	7.02	95.11

Table A21

Classification Accuracy of Guttman Errors across Simulation Conditions for Careless Responding

Dimension	**NPV**	**PPV**	**Sensitivity**	**Specificity**
1	0.93	0.60	30.61	97.78
2	0.93	0.59	30.08	97.70
3	0.92	0.51	25.90	97.23
4	0.92	0.55	28.12	97.47
Response Categories	**NPV**	**PPV**	**Sensitivity**	**Specificity**
4	0.94	0.80	40.59	98.88
5	0.94	0.83	42.06	99.04
6	0.92	0.55	28.23	97.48
7	0.90	0.07	3.82	94.77
Test Length	**NPV**	**PPV**	**Sensitivity**	**Specificity**
12	0.92	0.45	23.20	96.92
36	0.93	0.67	34.15	98.17

Table A22

Classification Accuracy of H^T_i across Simulation Conditions for Careless Responding

Dimension	**NPV**	**PPV**	**Sensitivity**	**Specificity**
1	0.92	0.39	20.06	96.56
2	0.90	0.18	9.06	95.33
3	0.90	0.18	9.10	95.39
4	0.91	0.25	12.68	95.81
Response Categories	**NPV**	**PPV**	**Sensitivity**	**Specificity**
4	0.92	0.39	19.75	96.55
5	0.91	0.33	16.54	96.21
6	0.91	0.20	10.16	95.49
7	0.90	0.09	4.44	94.84
Test Length	**NPV**	**PPV**	**Sensitivity**	**Specificity**
12	0.90	0.18	9.29	95.40
36	0.91	0.32	16.16	96.15

Table A23

Classification Accuracy of U3 across Simulation Conditions for Careless Responding

Dimension	**NPV**	**PPV**	**Sensitivity**	**Specificity**
1	0.92	0.55	28.09	97.48
2	0.92	0.51	25.63	97.23
3	0.92	0.46	23.32	96.96
4	0.92	0.48	24.17	97.05
Response Categories	**NPV**	**PPV**	**Sensitivity**	**Specificity**
4	0.93	0.74	37.75	98.55
5	0.94	0.77	39.16	98.73
6	0.92	0.47	23.98	97.02
7	0.90	0.01	0.31	94.42
Test Length	**NPV**	**PPV**	**Sensitivity**	**Specificity**
12	0.92	0.39	20.06	96.60
36	0.93	0.60	30.54	97.76

Table A24

Aggregated Results for Guttman Errors across Simulation Conditions

Aberrancy	NPV	PPV	Sensitivity	Specificity
Acquiescence	0.90	0.10	0.05	0.95
Disacquiescence	0.91	0.24	0.13	0.96
Midpoint	0.90	0.00	0.00	0.94
Extreme	0.92	0.52	0.27	0.97
Social	0.90	0.15	0.08	0.95
Careless	0.93	0.57	0.29	0.98
MEAN	**0.91**	**0.26**	**0.13**	**0.96**

Table A25

Aggregated Results for H^T_i across Simulation Conditions

Aberrancy	NPV	PPV	Sensitivity	Specificity
Acquiescence	0.90	0.16	0.08	0.95
Disacquiescence	0.90	0.08	0.04	0.95
Midpoint	0.91	0.25	0.13	0.96
Extreme	0.90	0.10	0.05	0.95
Social	0.90	0.18	0.09	0.95
Careless	0.91	0.26	0.13	0.96
MEAN	**0.90**	**0.17**	**0.09**	**0.95**

Table A26

Aggregated Results for U3 across Simulation Conditions

Aberrancy	NPV	PPV	Sensitivity	Specificity
Acquiescence	0.90	0.08	0.05	0.95
Disacquiescence	0.91	0.24	0.12	0.96
Midpoint	0.90	0.00	0.00	0.94
Extreme	0.92	0.48	0.25	0.97
Social	0.90	0.13	0.07	0.95
Careless	0.92	0.49	0.25	0.97
MEAN	**0.91**	**0.24**	**0.12**	**0.96**

APPENDIX B

Figure B1

General Factor Structure of the Simulated Data

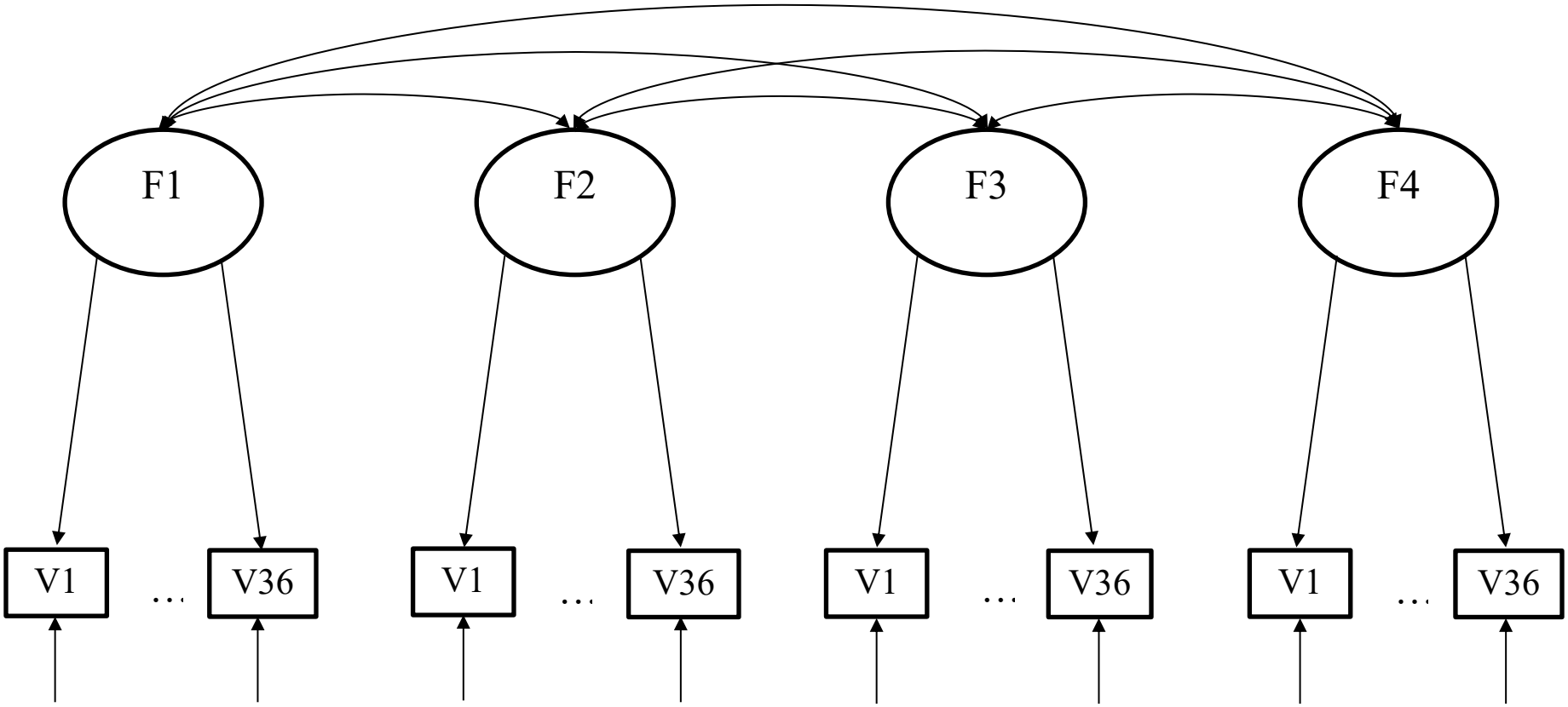


Figure B2

NPV of the Number of Normed Guttman Errors by Test Length when Applied to Acquiescence Responding

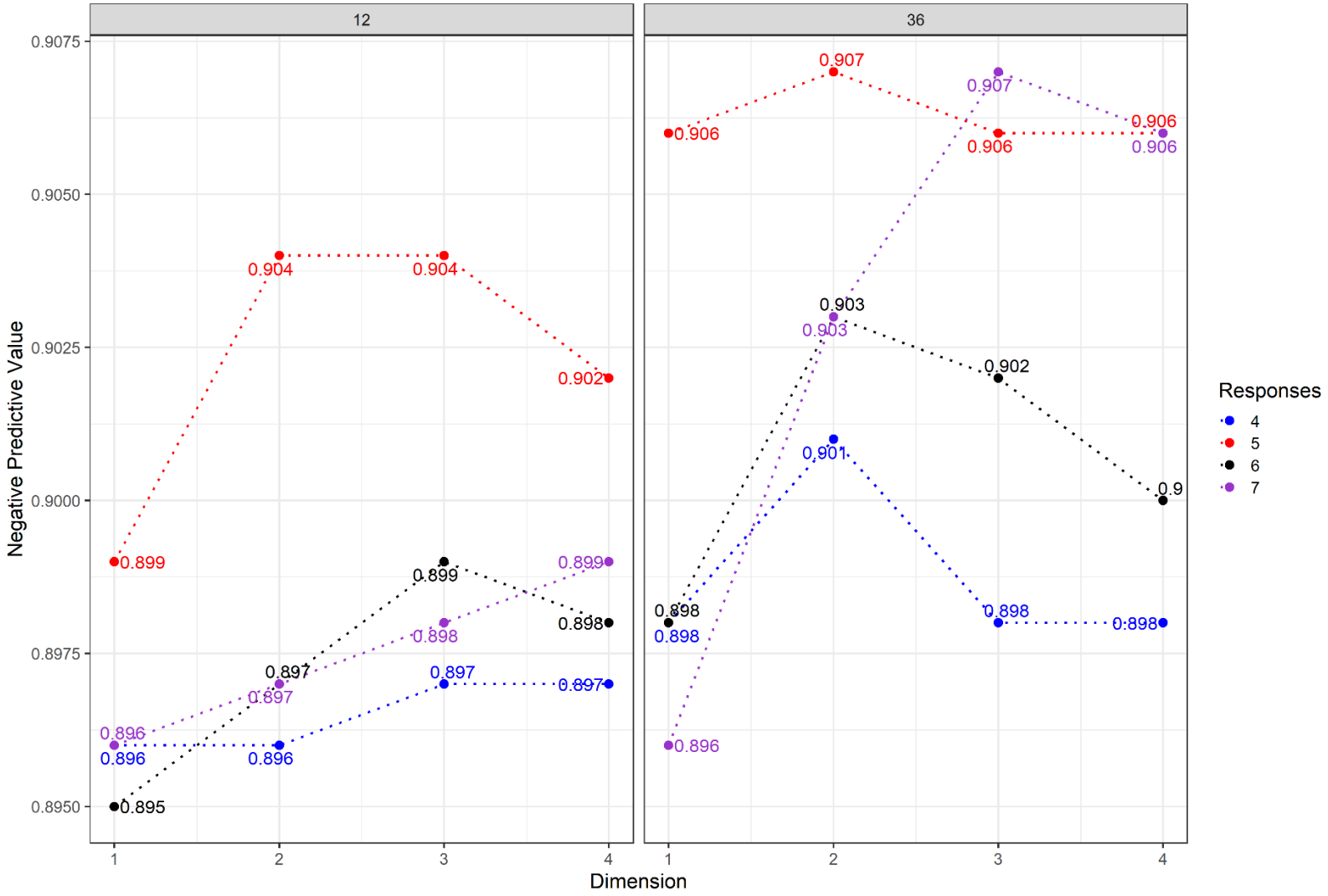


Figure B3

PPV of the Number of Normed Guttman Errors by Test Length when Applied to Acquiescence Responding

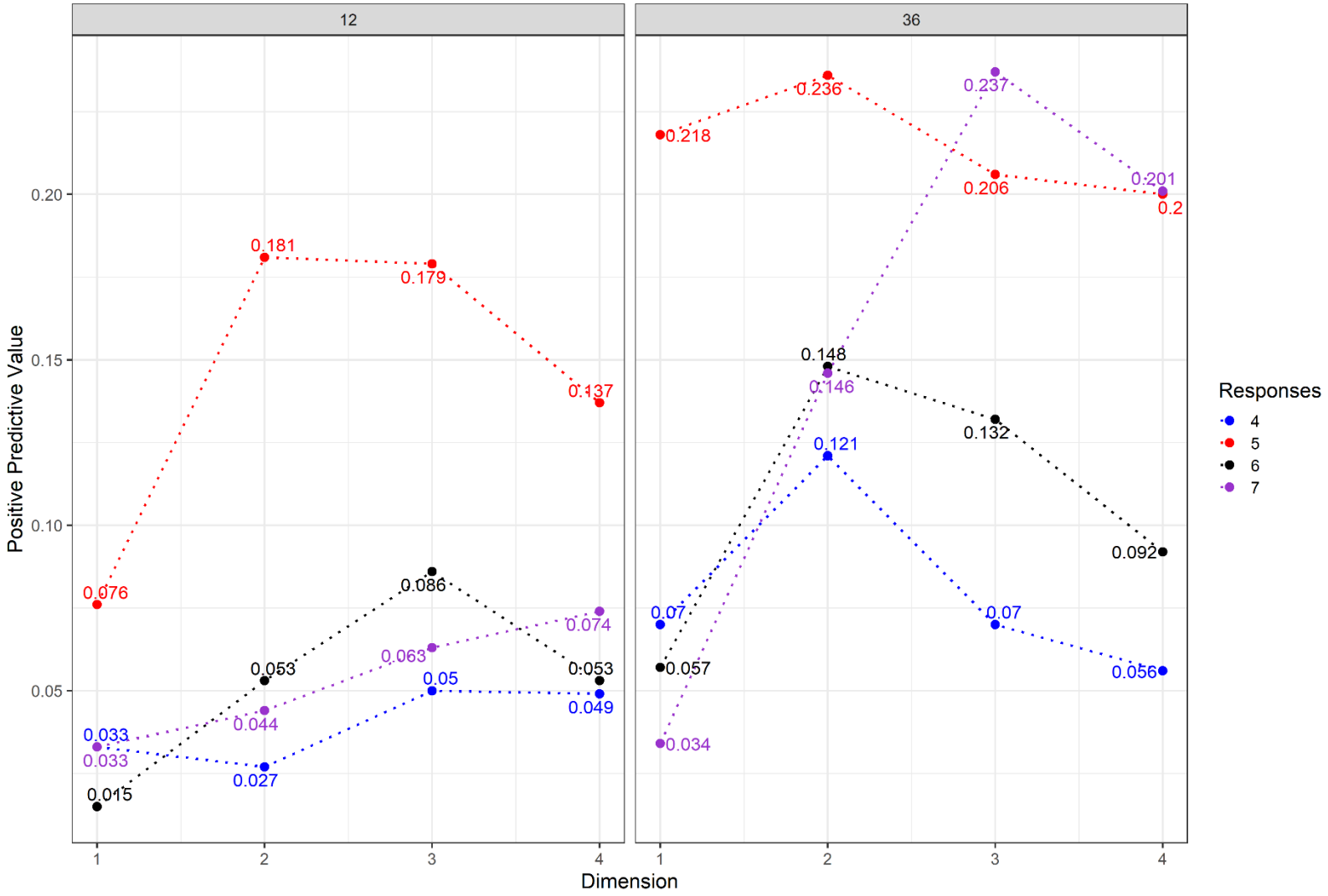


Figure B4

Sensitivity of the Number of Normed Guttman Errors by Test Length when Applied to Acquiescence Responding

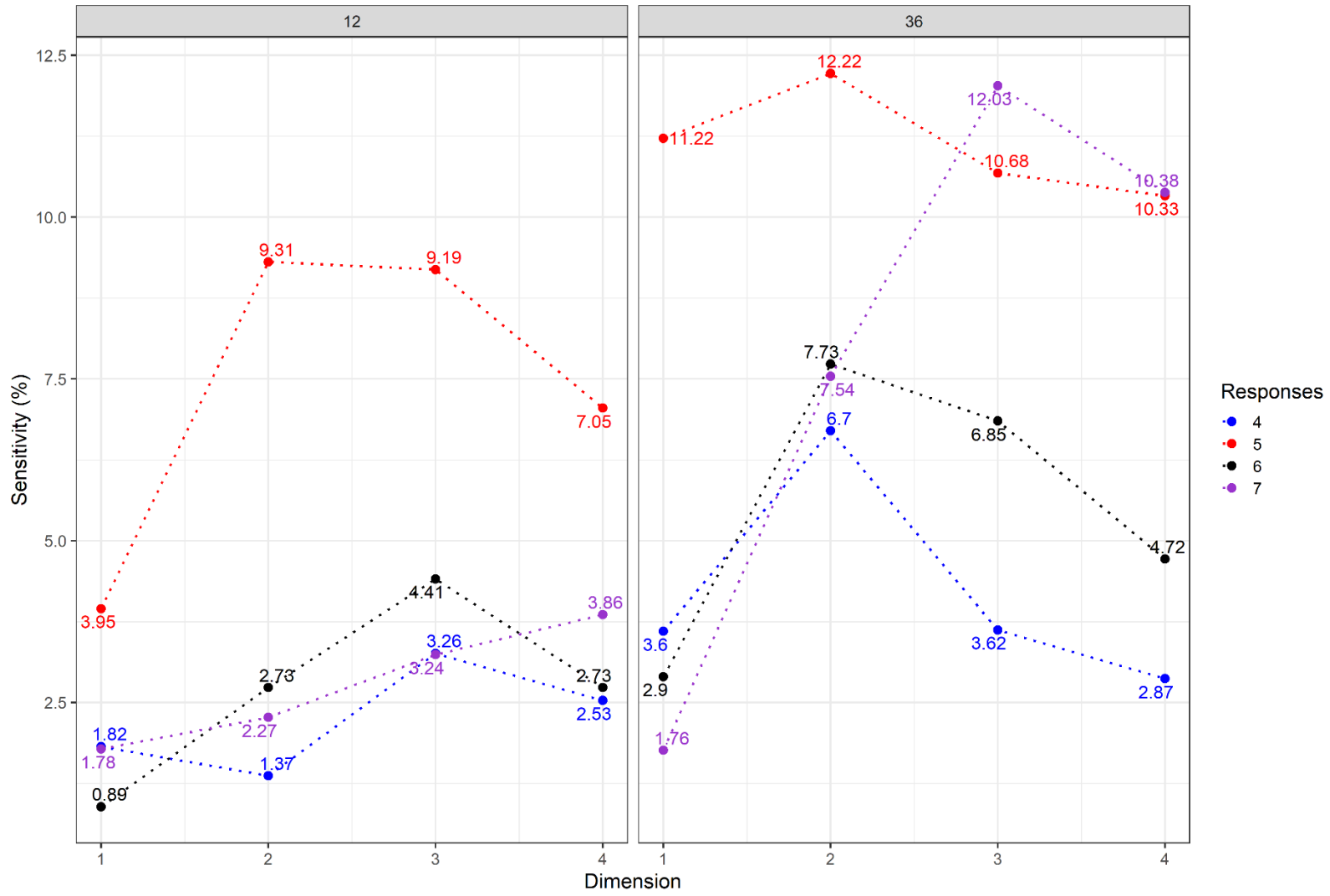


Figure B5

Specificity of the Number of Normed Guttman Errors by Test Length when Applied to Acquiescence Responding

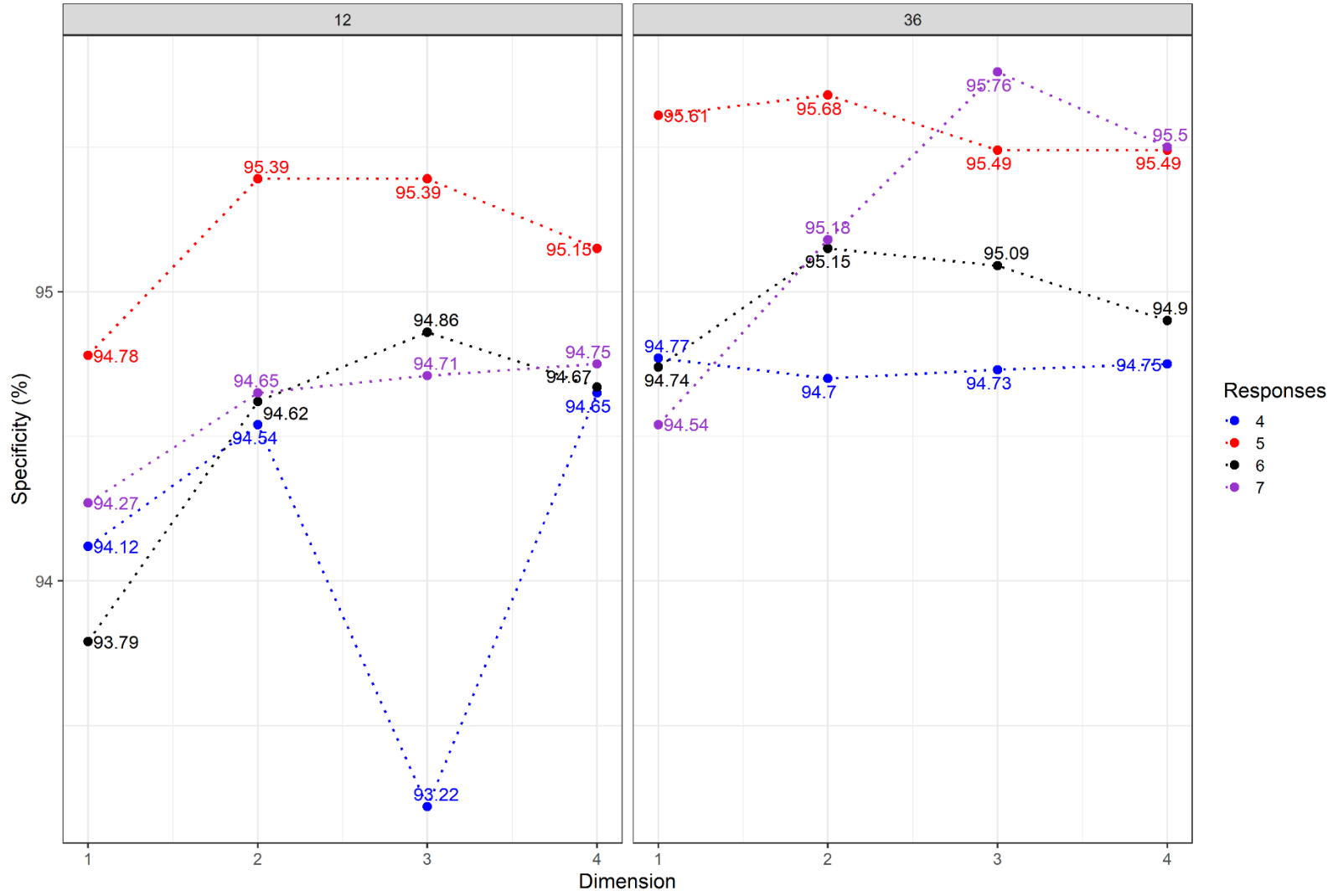


Figure B6

Negative Predictive Value of H^T_i by Test Length when Applied to Acquiescence Responding

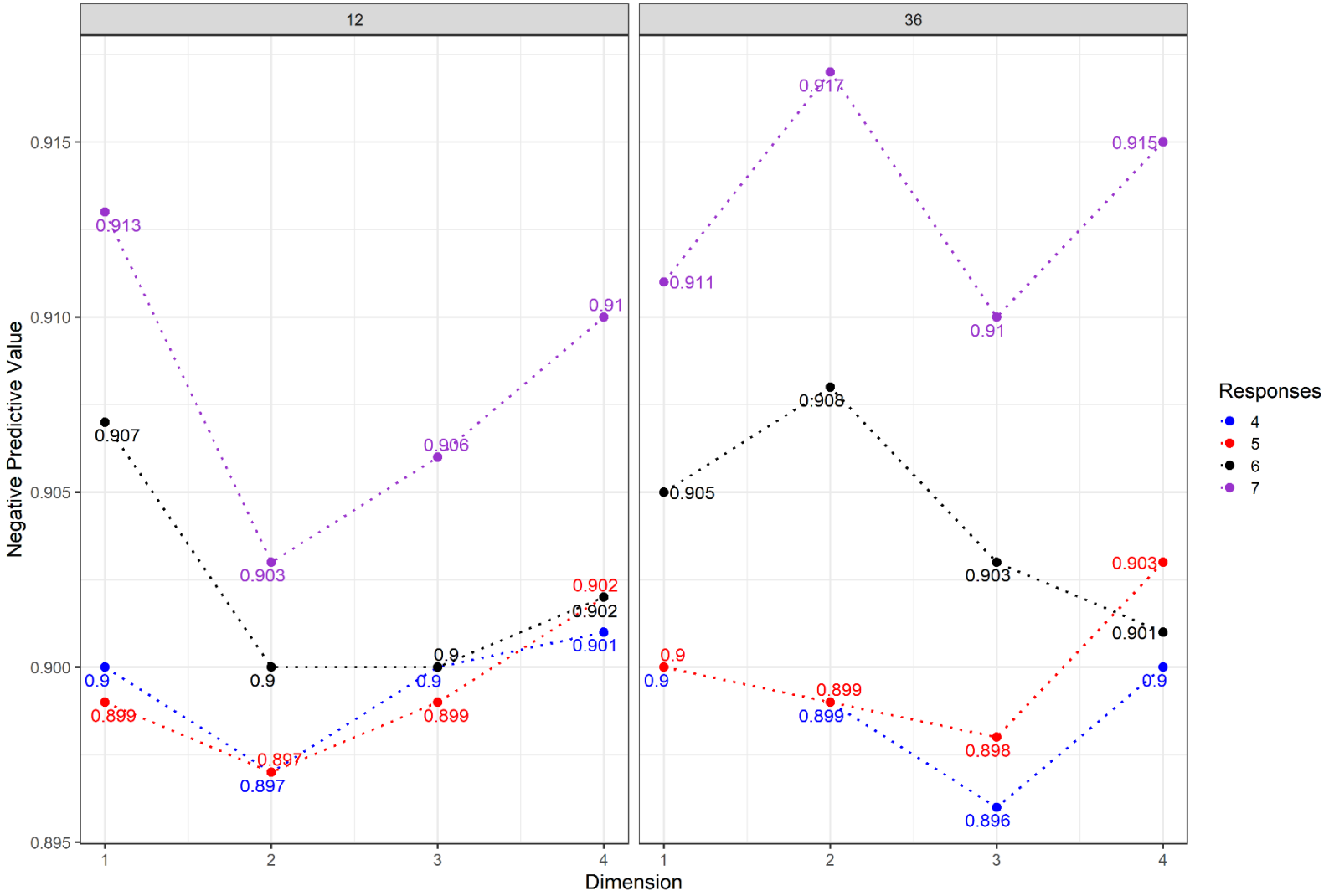


Figure B7

Positive Predictive Value of H^T_i by Test Length when Applied to Acquiescence Responding

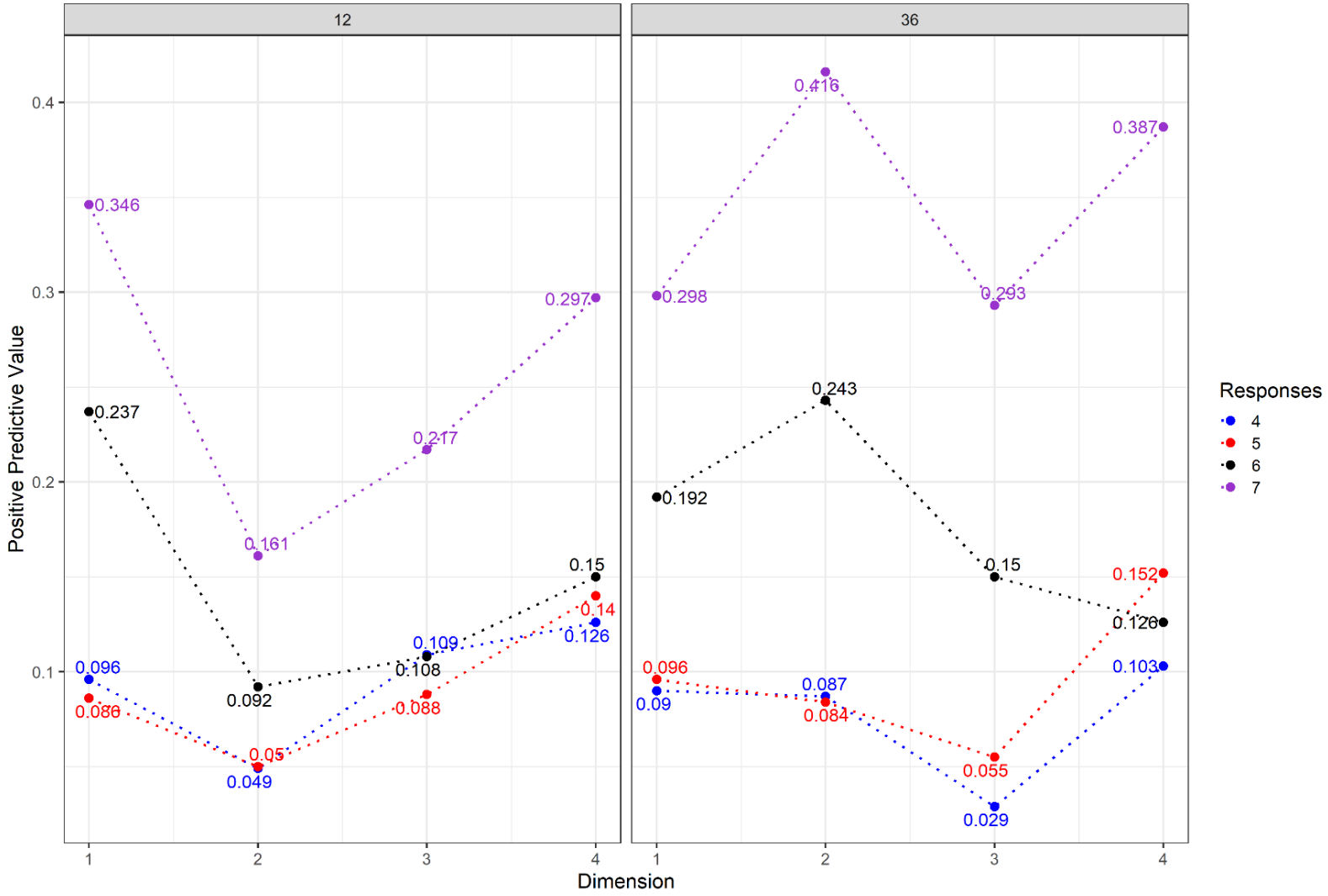


Figure B8

Sensitivity of H^T_i by Test Length when Applied to Acquiescence Responding

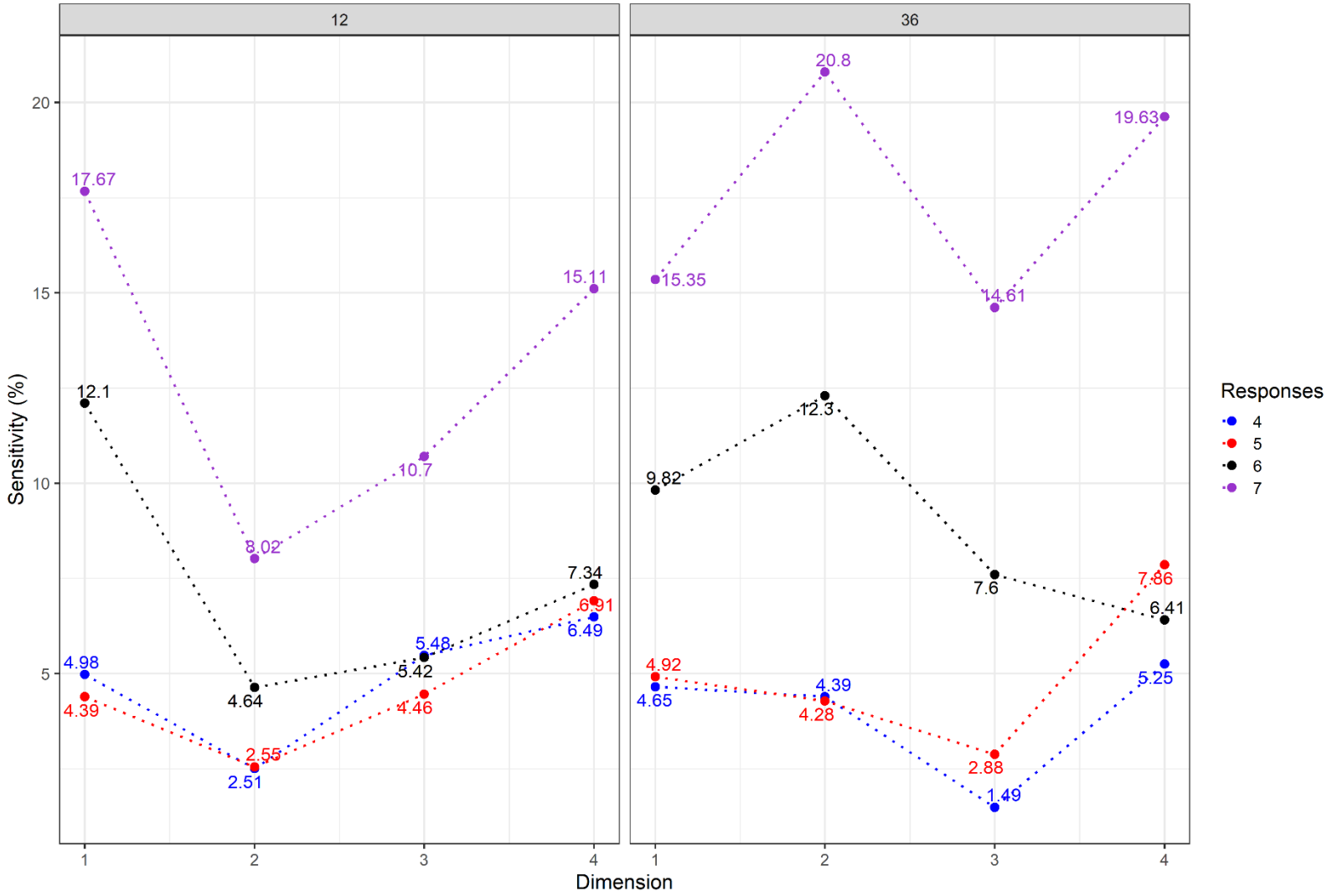


Figure B9

Specificity of H^T_i by Test Length when Applied to Acquiescence Responding

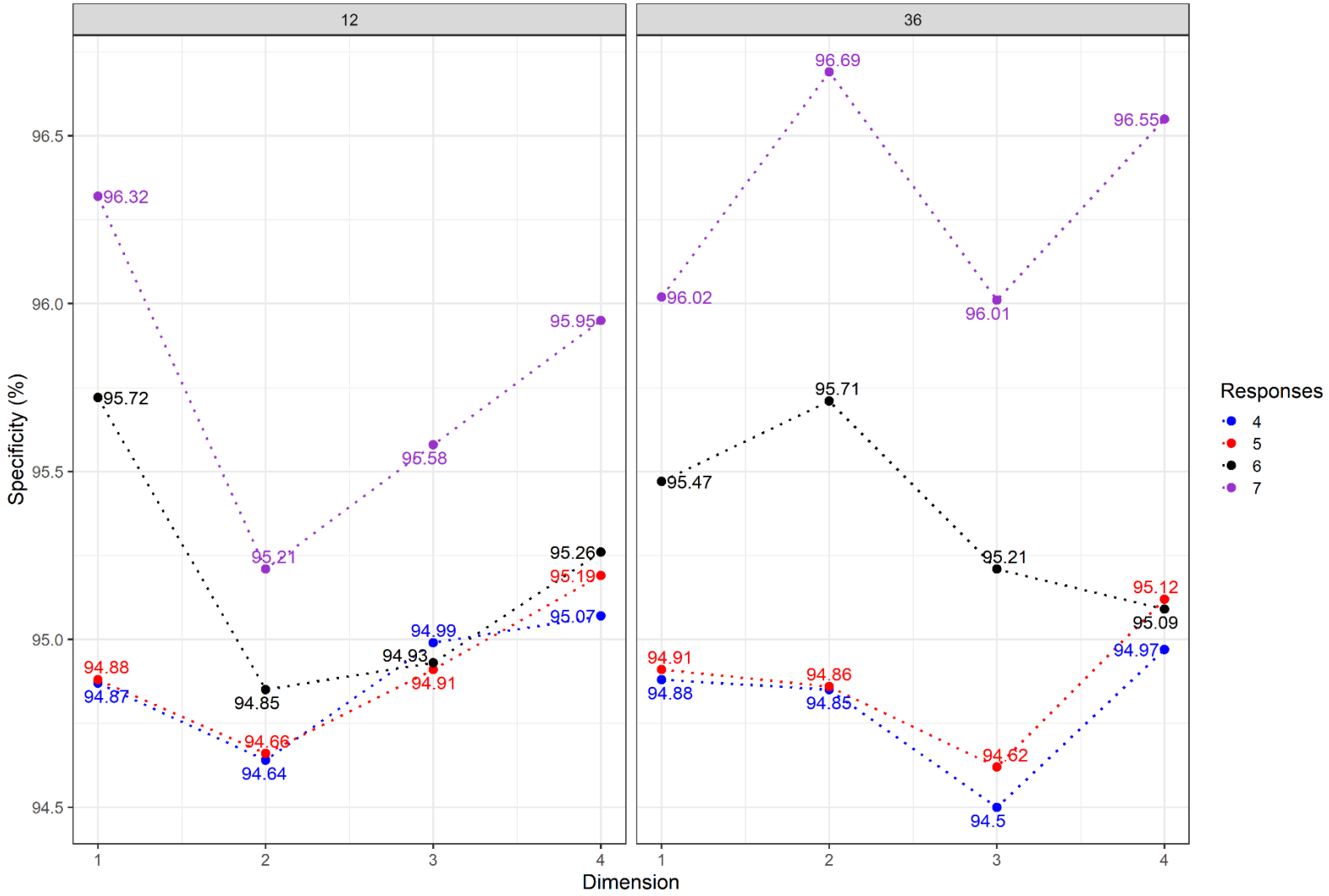


Figure B10

Negative Predictive Value of U3 by Test Length when Applied to Acquiescence Responding

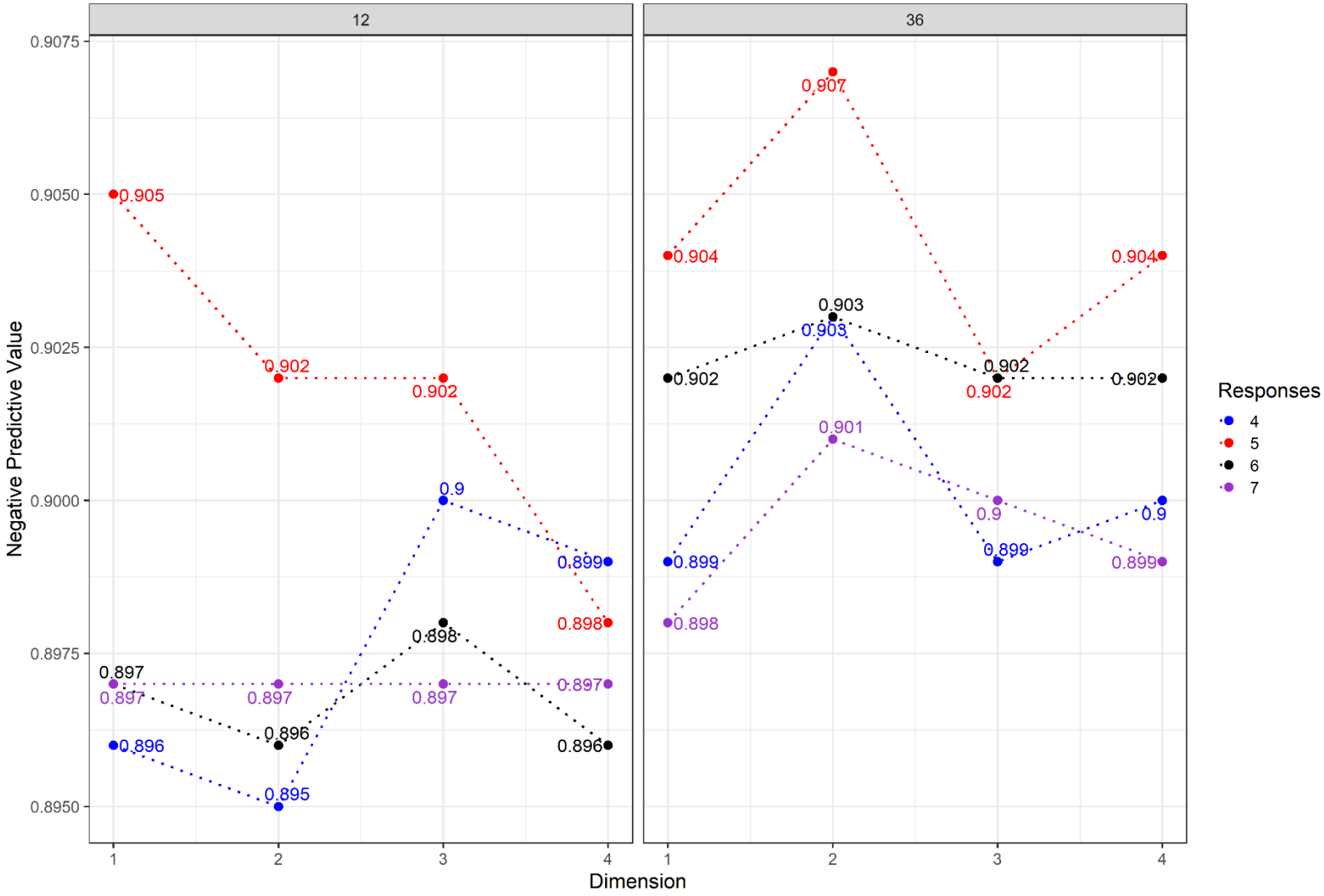


Figure B11

Positive Predictive Value of U3 by Test Length when Applied to Acquiescence Responding

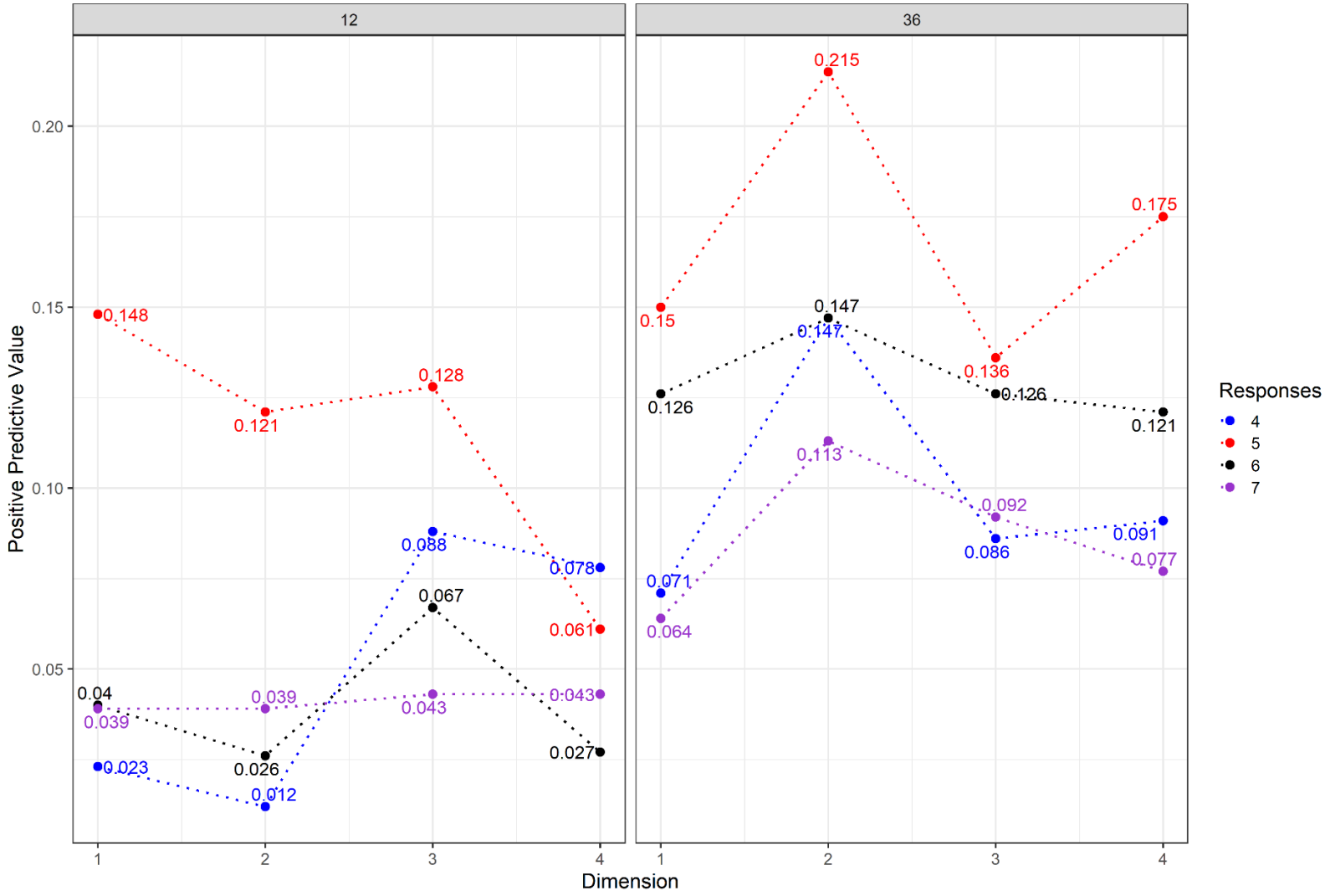


Figure B12

Sensitivity of U3 by Test Length when Applied to Acquiescence Responding

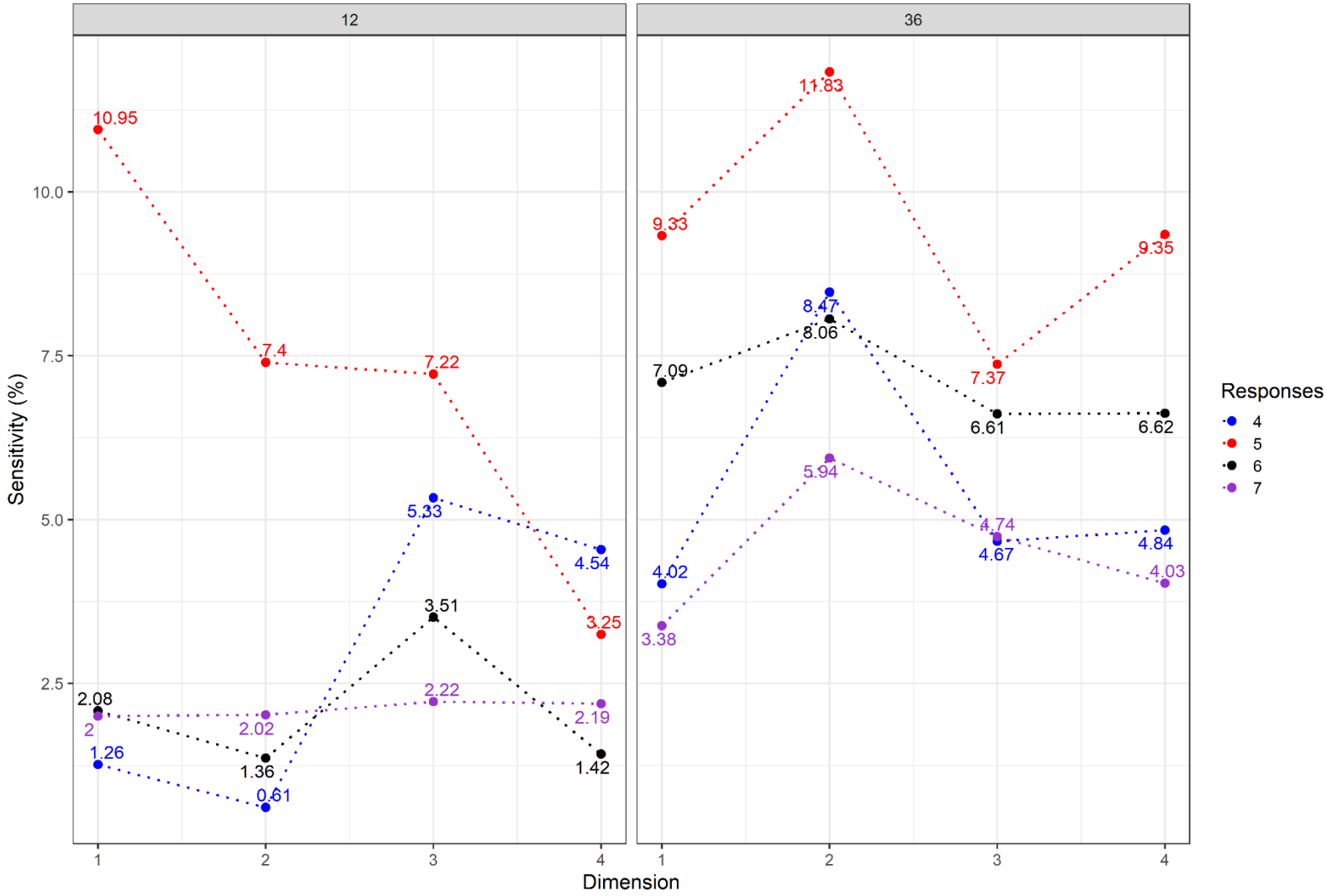


Figure B13

Specificity of U3 by Test Length when Applied to Acquiescence Responding

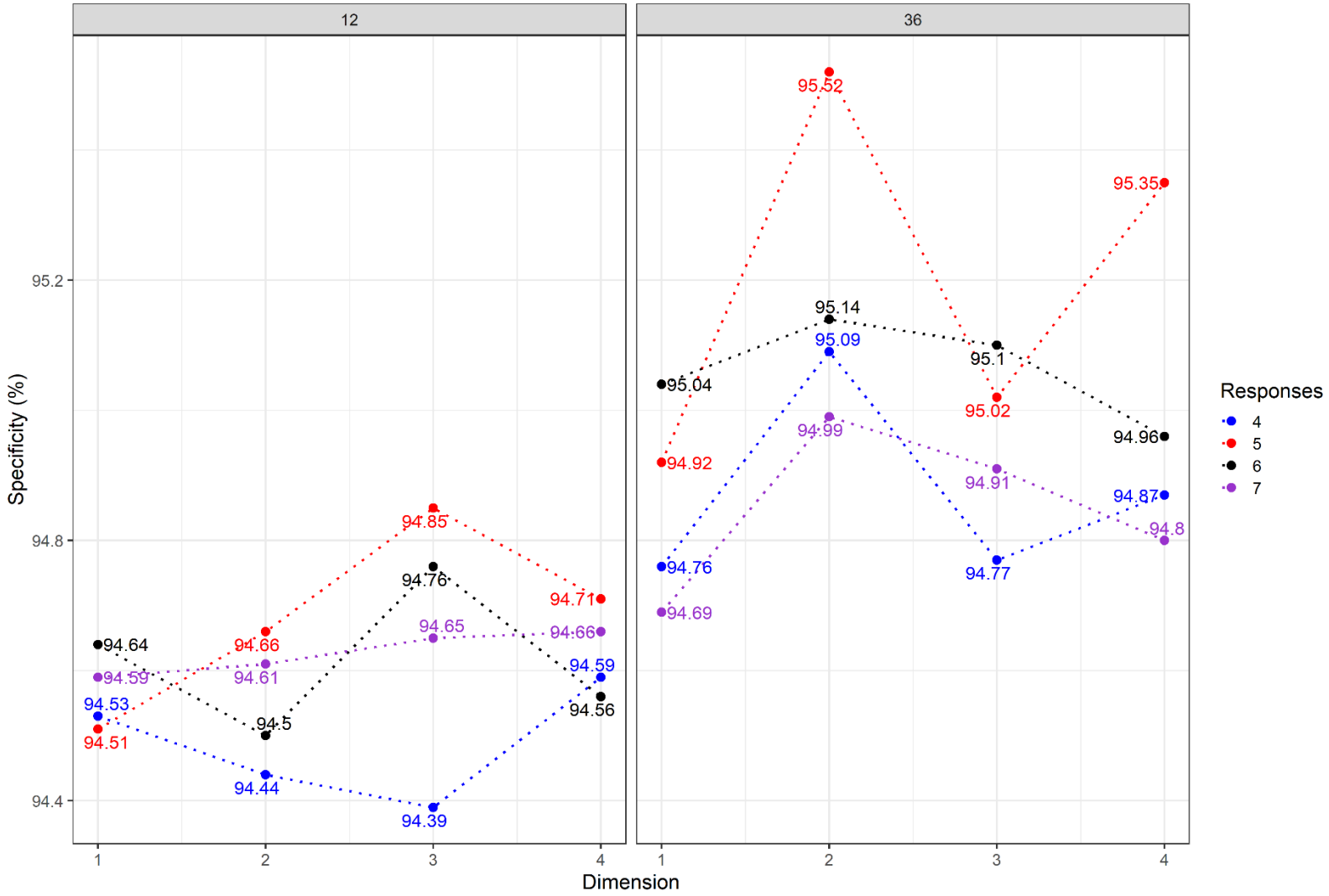


Figure B14

Negative Predictive Value of Guttman Errors by Test Length when Applied to Disacquiescence Responding

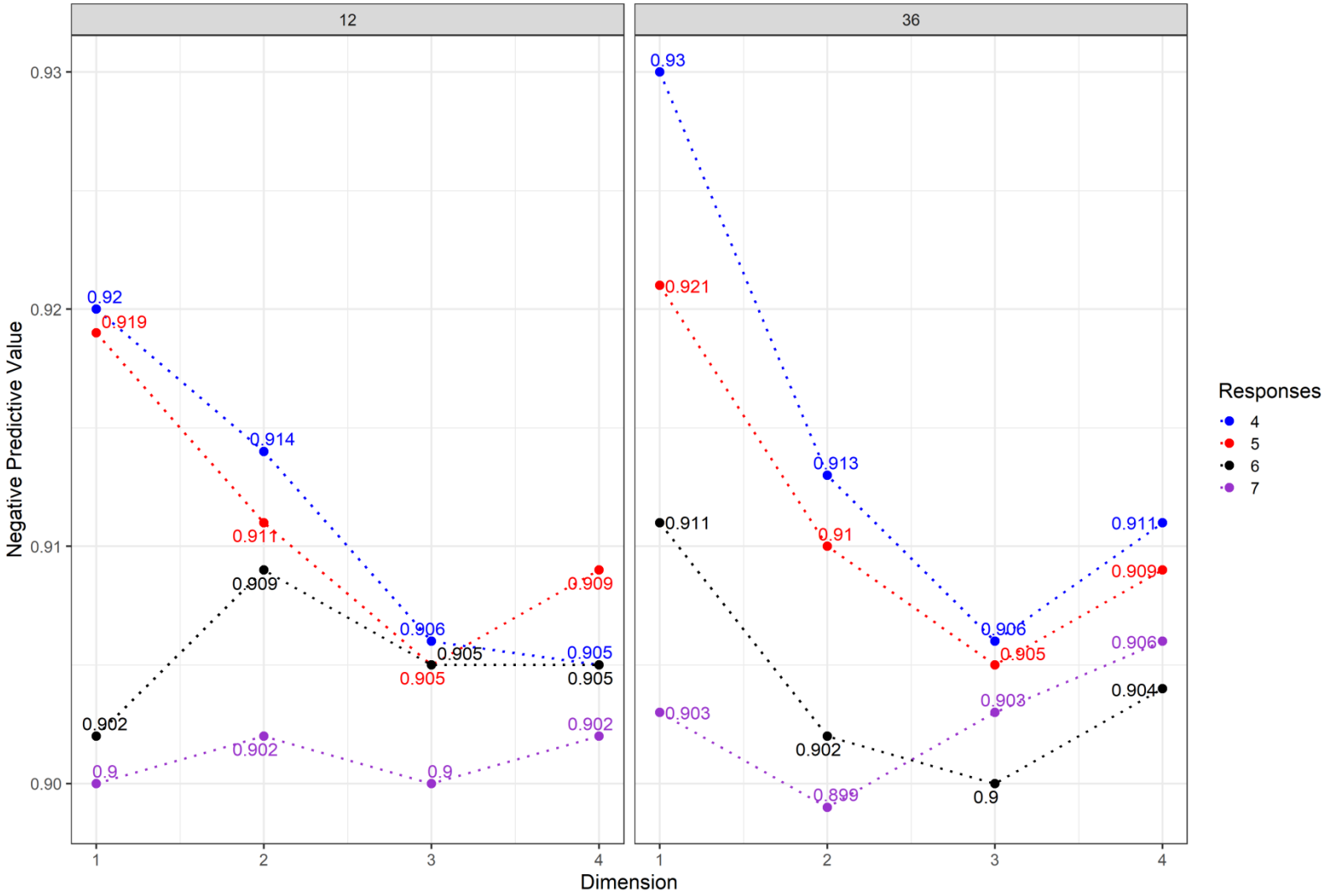


Figure B15

Positive Predictive Value of Guttman Errors by Test Length when Applied to Disacquiescence Responding

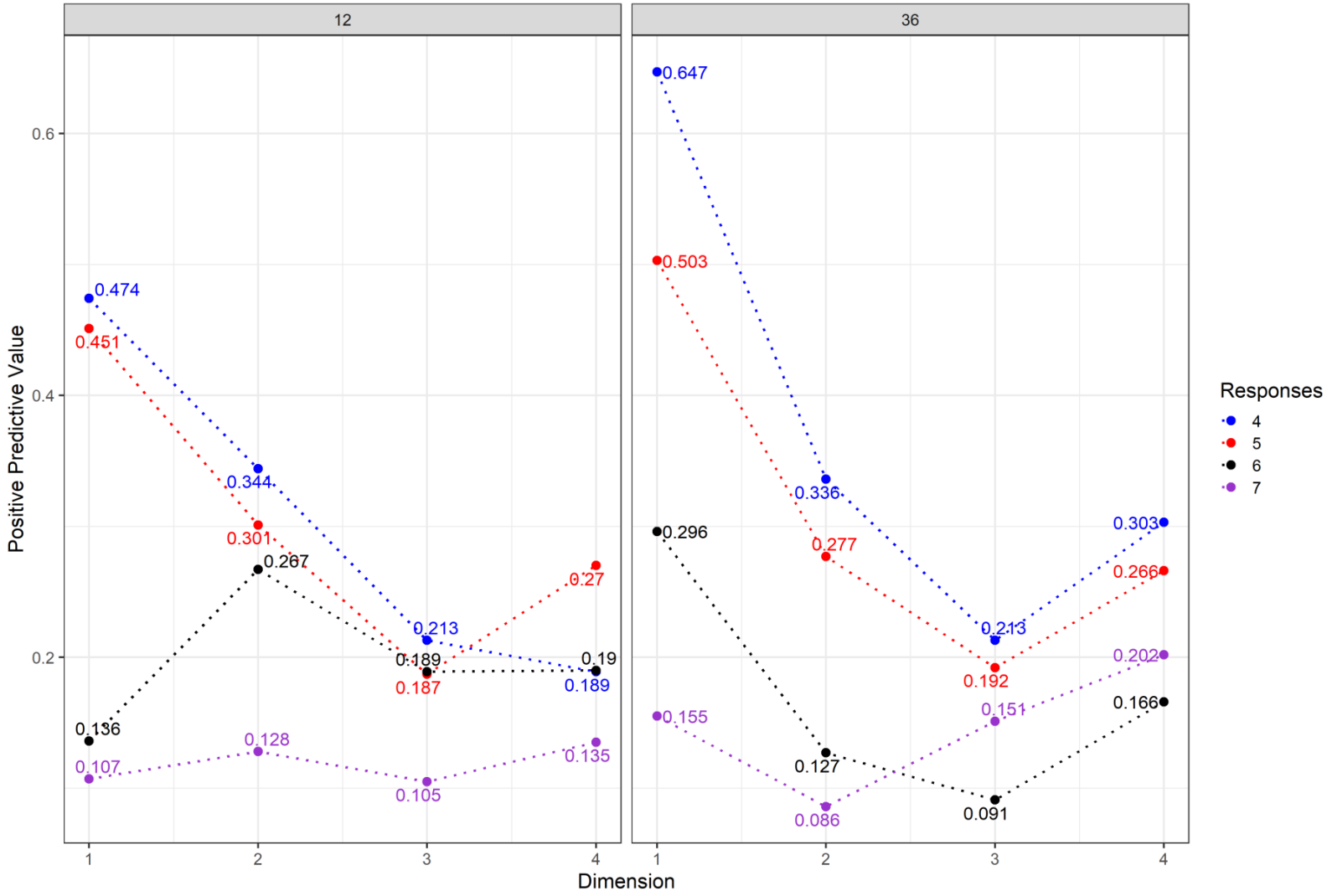


Figure B16

Sensitivity of Guttman Errors by Test Length when Applied to Disacquiescence Responding

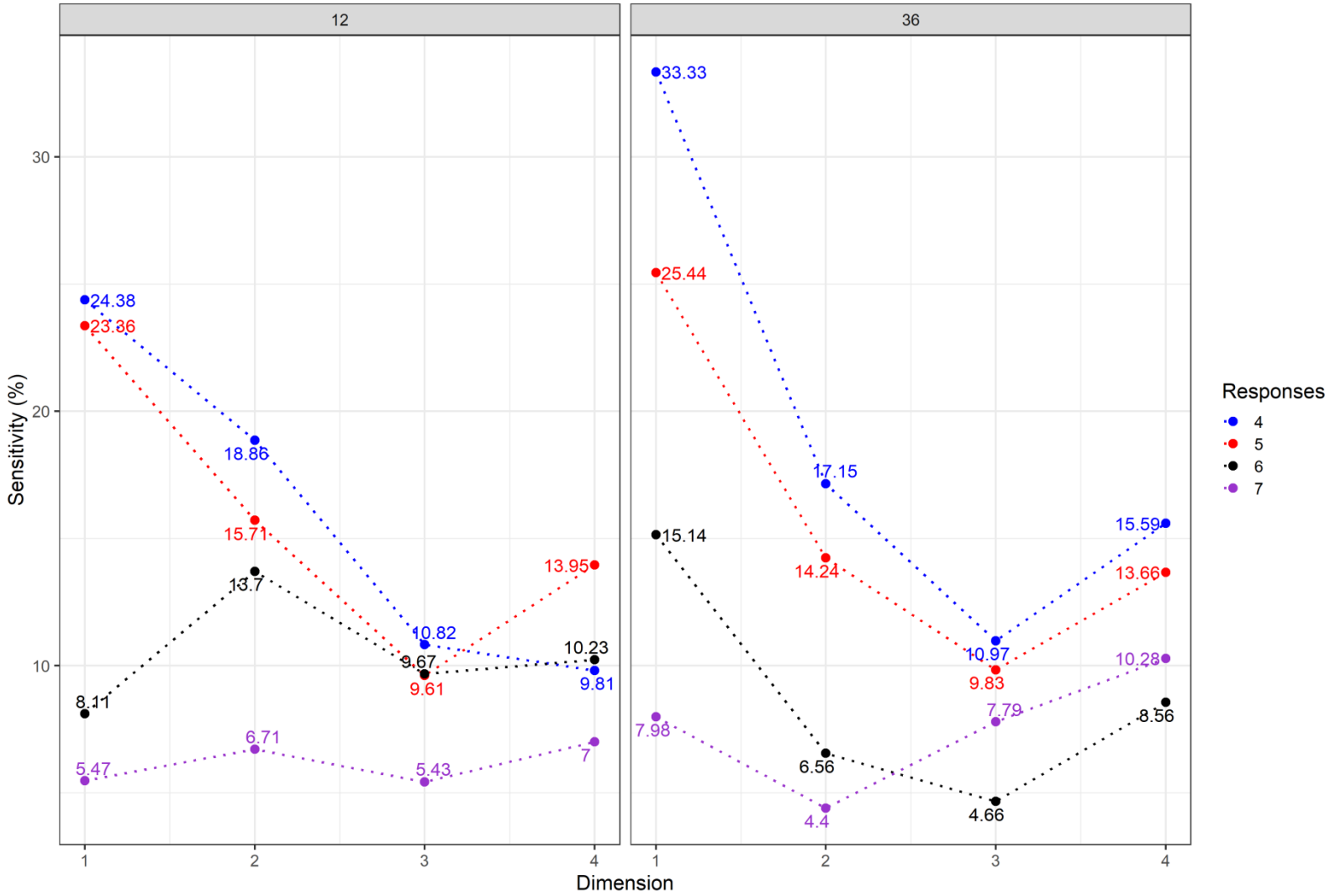


Figure B17

Specificity of Guttman Errors by Test Length when Applied to Disacquiescence Responding

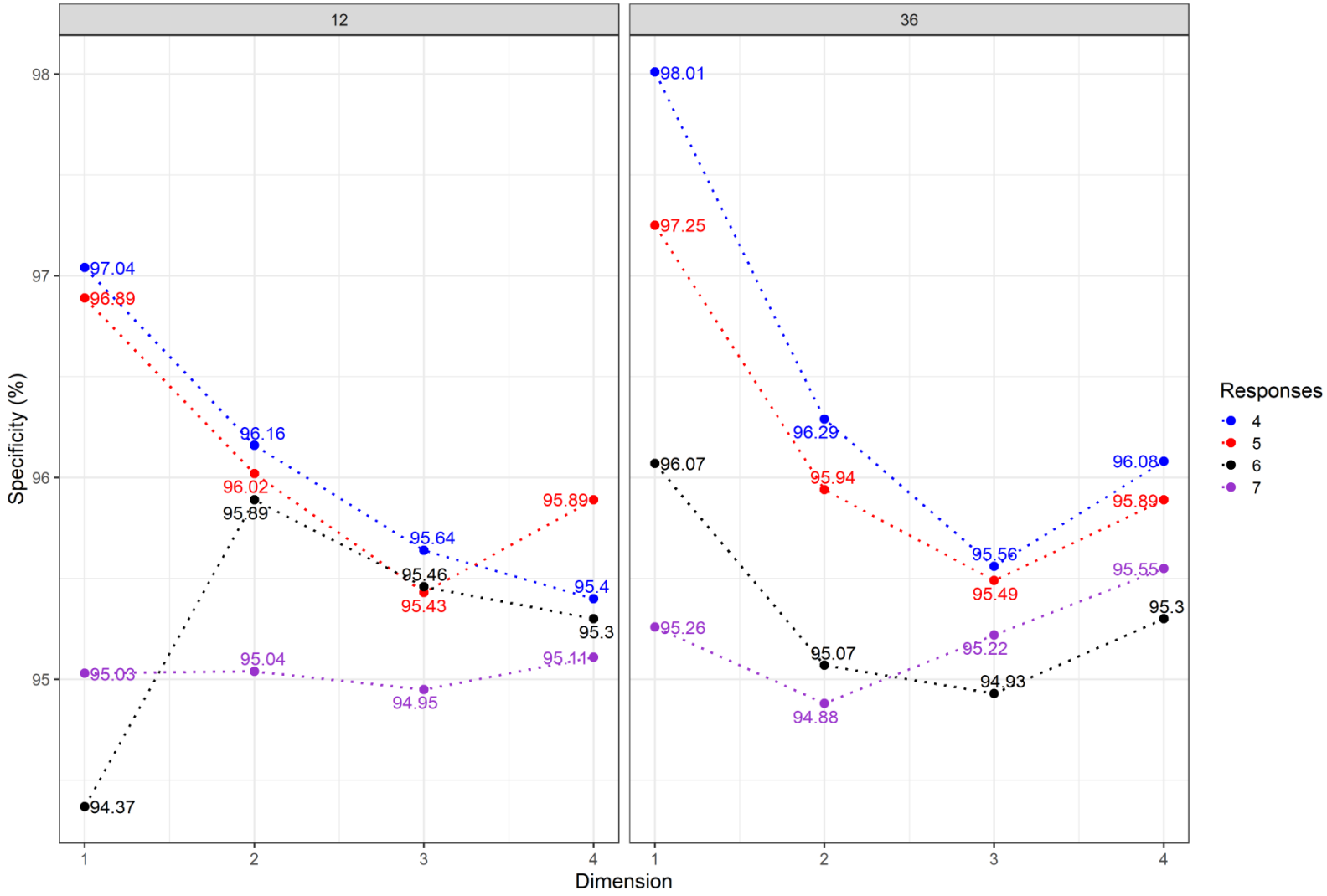


Figure B18

Negative Predictive Value of H^T_i by Test Length when Applied to Disacquiescence Responding

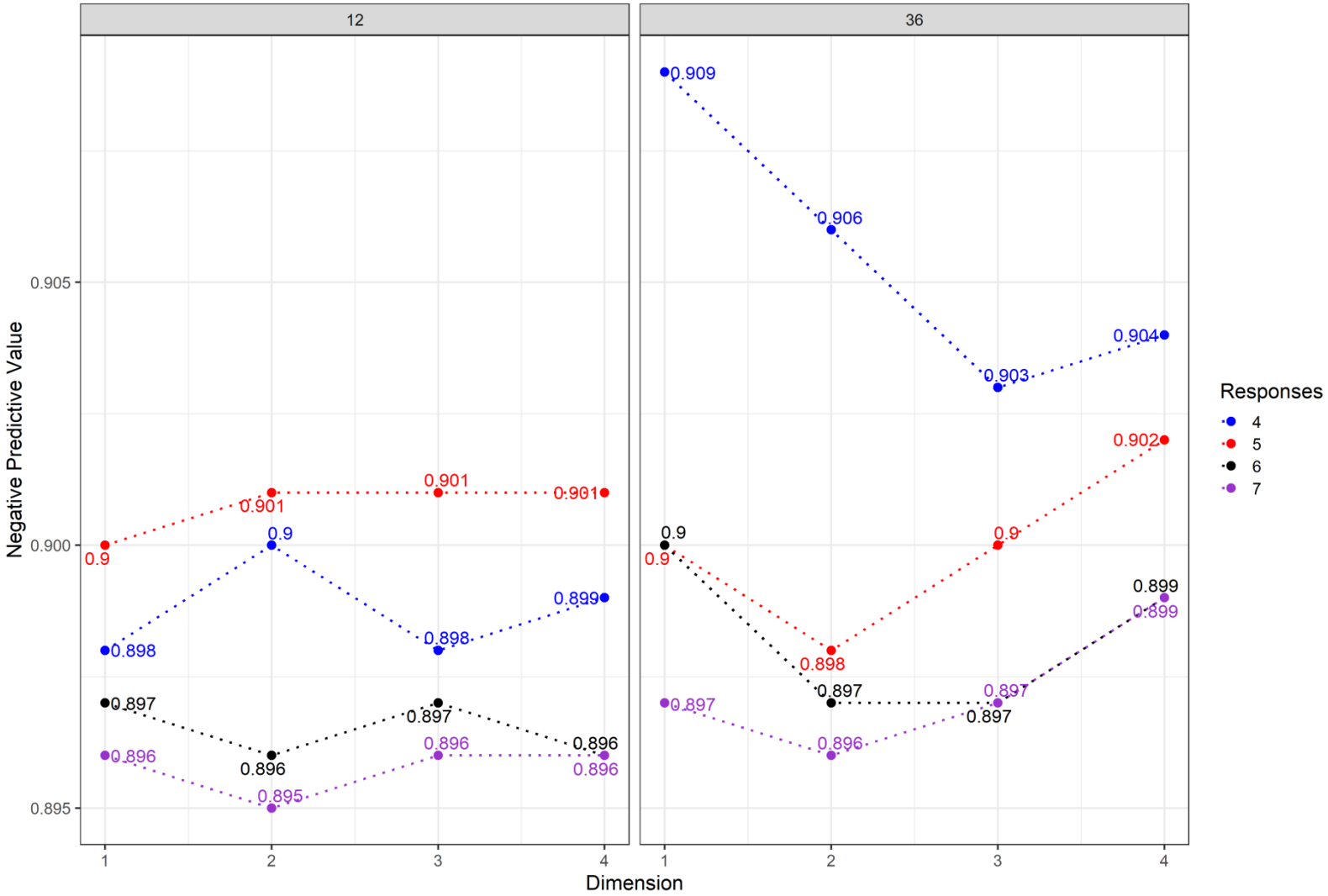


Figure B19

Positive Predictive Value of H^T_i by Test Length when Applied to Disacquiescence Responding

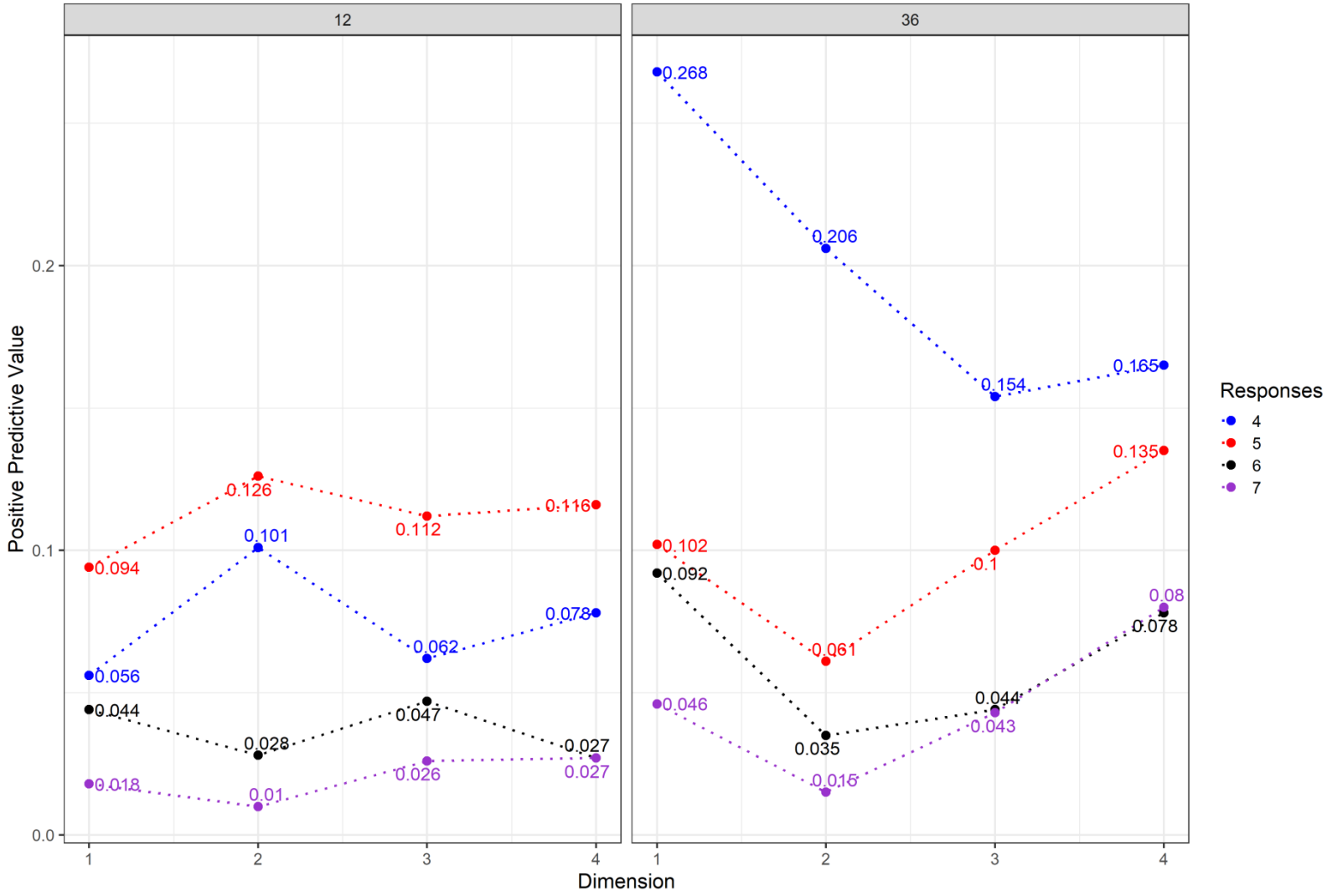


Figure B20

Sensitivity of H^T_i by Test Length when Applied to Disacquiescence Responding

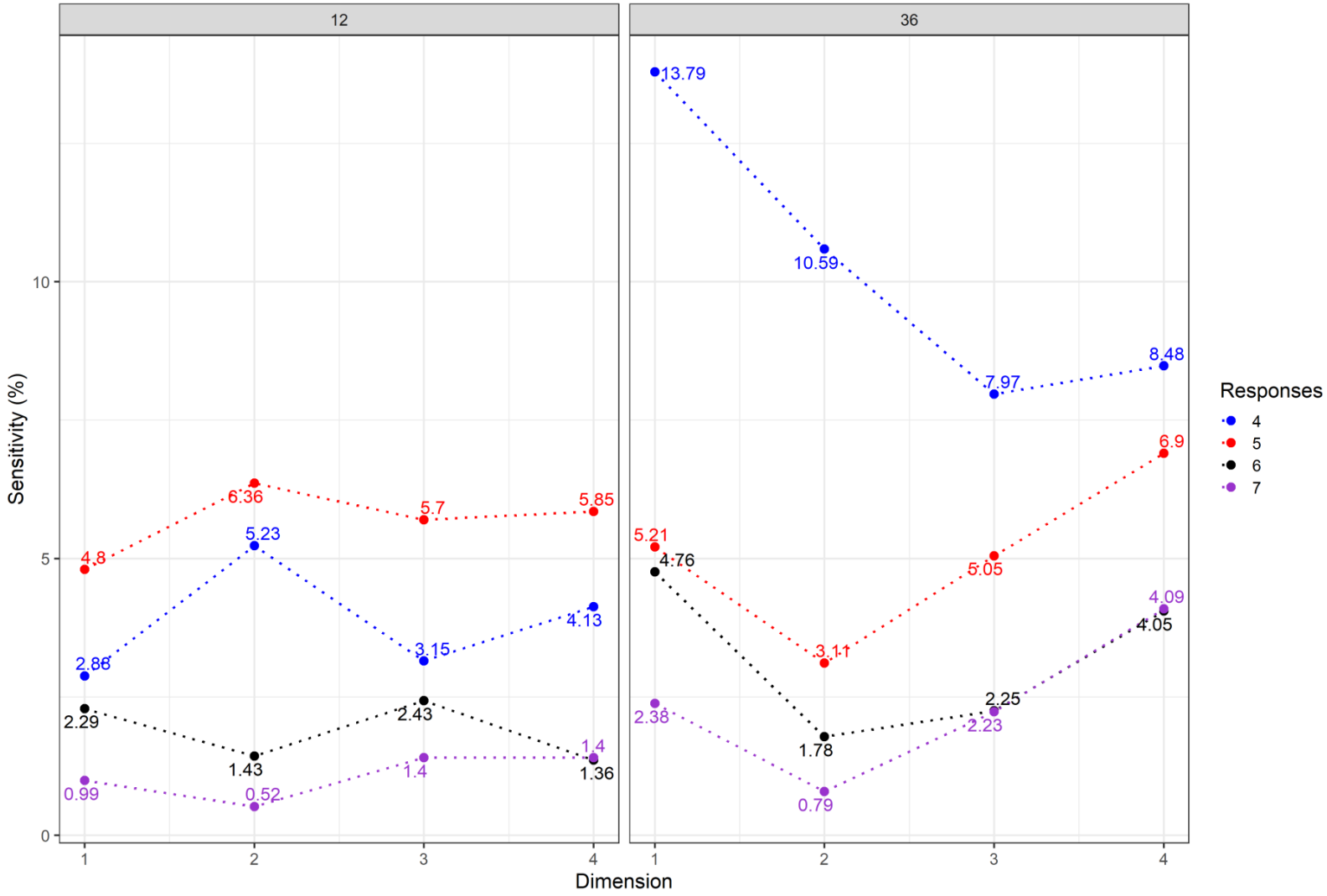


Figure B21

Specificity of H^T_i by Test Length when Applied to Disacquiescence Responding

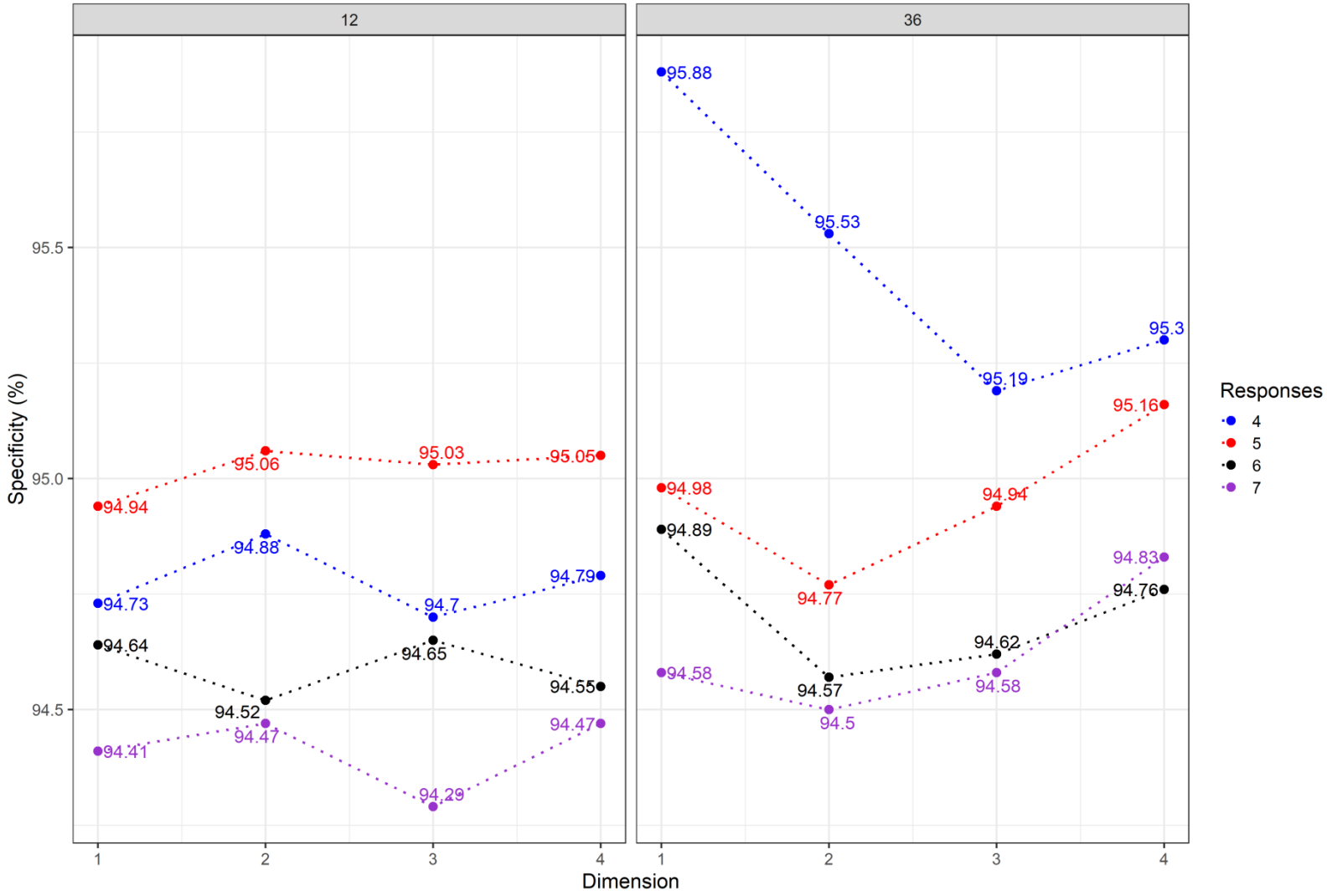


Figure B22

Negative Predictive Value of U3 by Test Length when Applied to Disacquiescence Responding

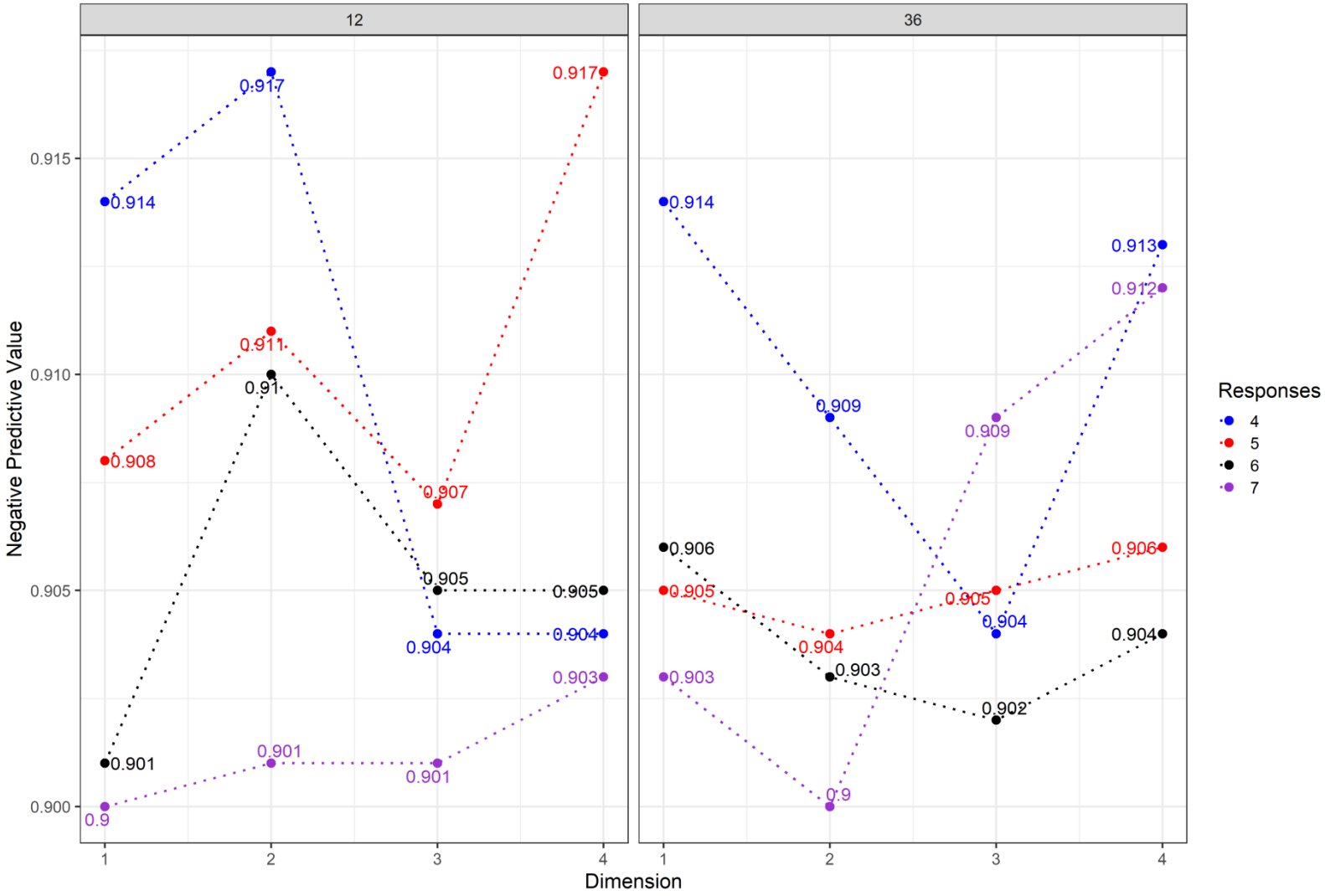


Figure B23

Positive Predictive Value of U3 by Test Length when Applied to Disacquiescence Responding

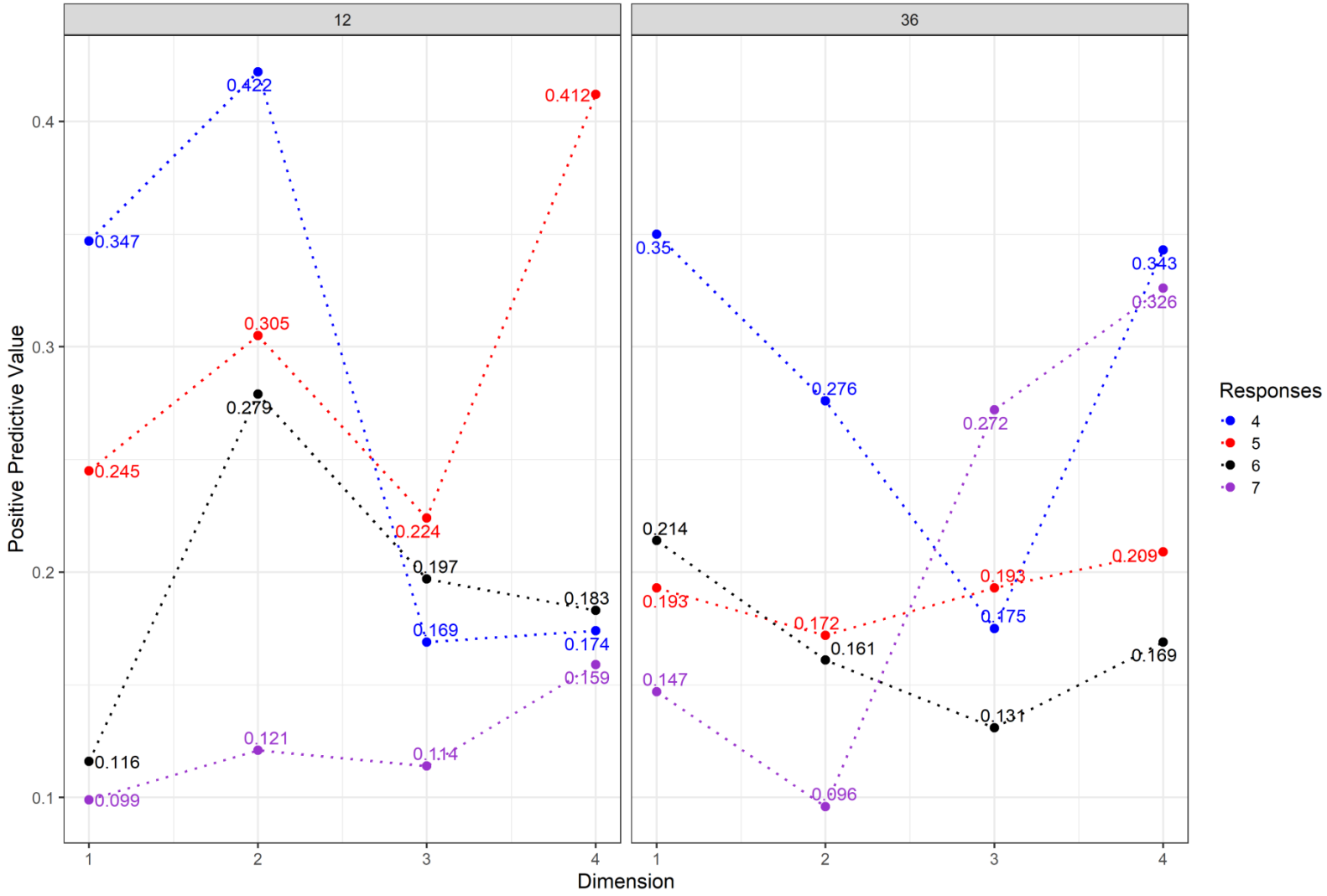


Figure B24

Sensitivity of U3 by Test Length when Applied to Disacquiescence Responding

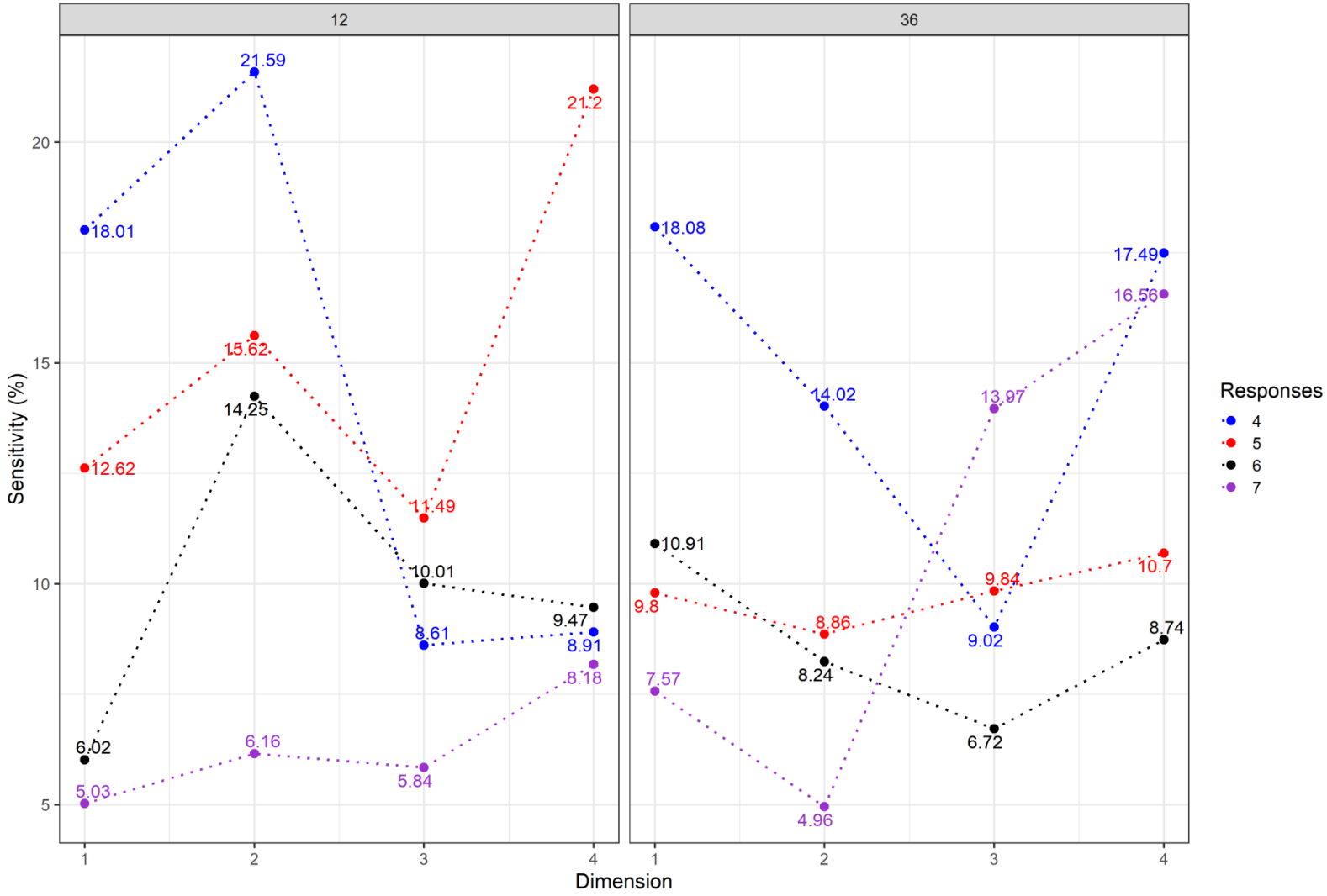


Figure B25

Specificity of U3 by Test Length when Applied to Disacquiescence Responding

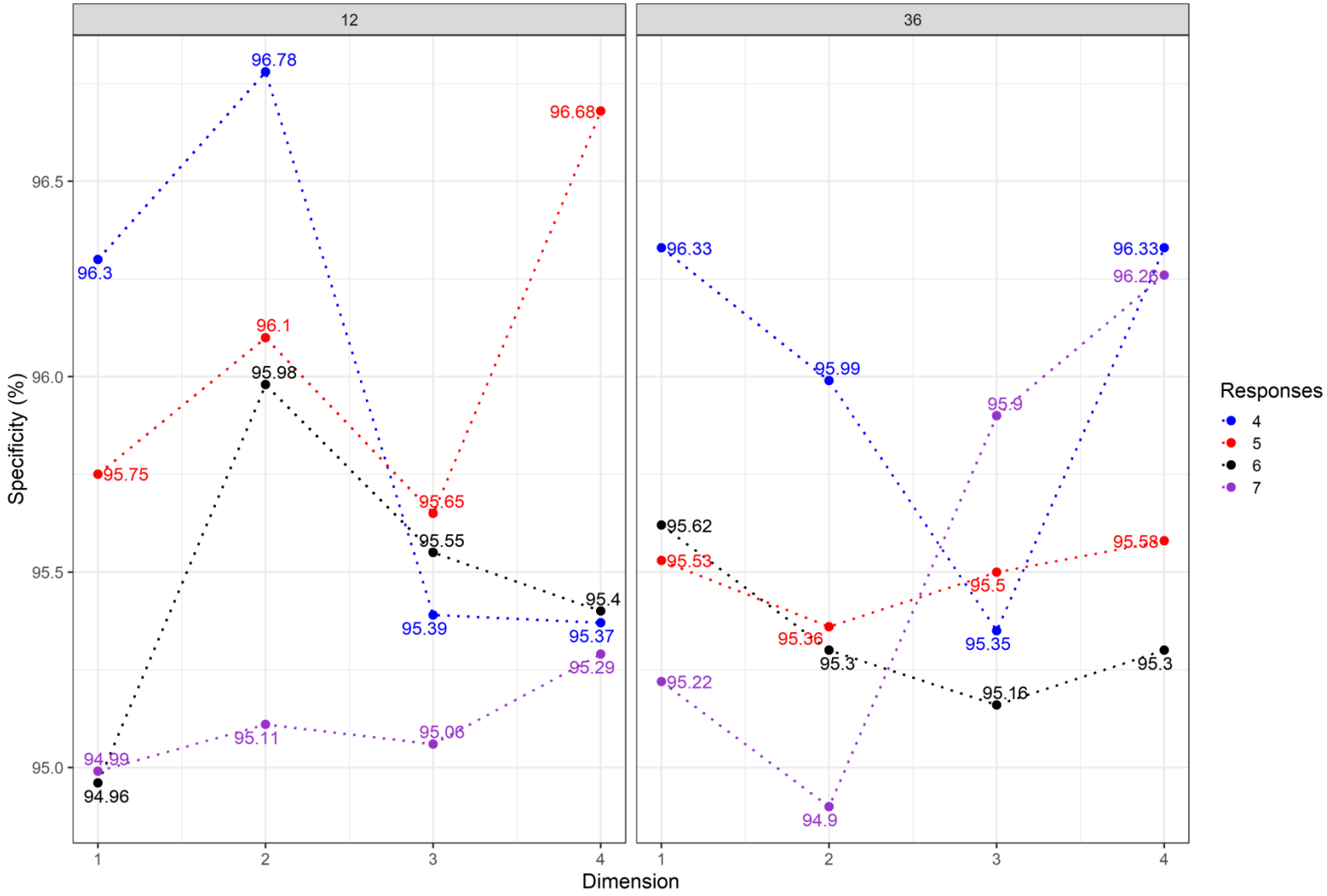


Figure B26

Negative Predictive Value of Guttman Errors by Test Length when Applied to Midpoint Responding

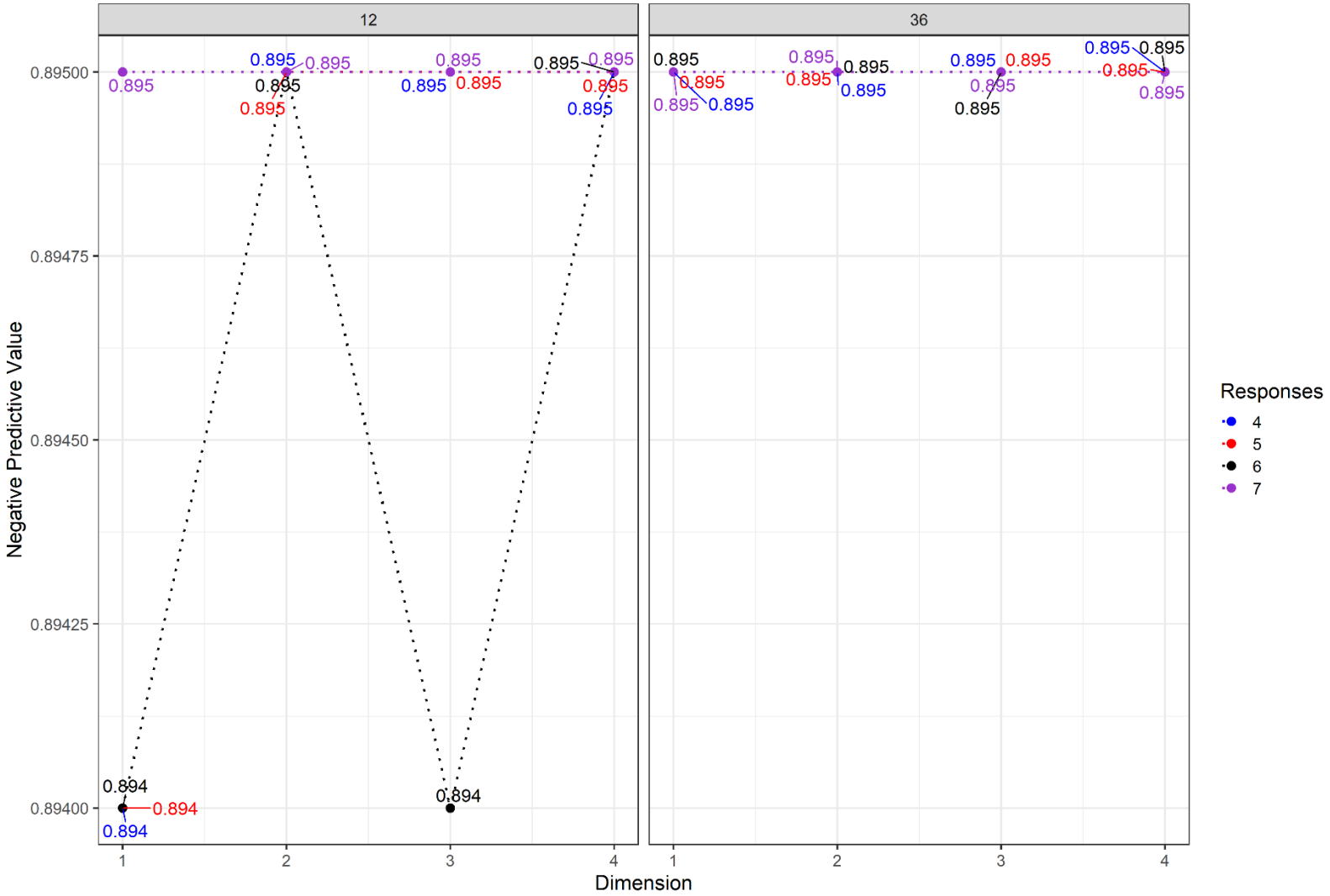


Figure B27

Positive Predictive Value of Guttman Errors by Test Length when Applied to Midpoint Responding

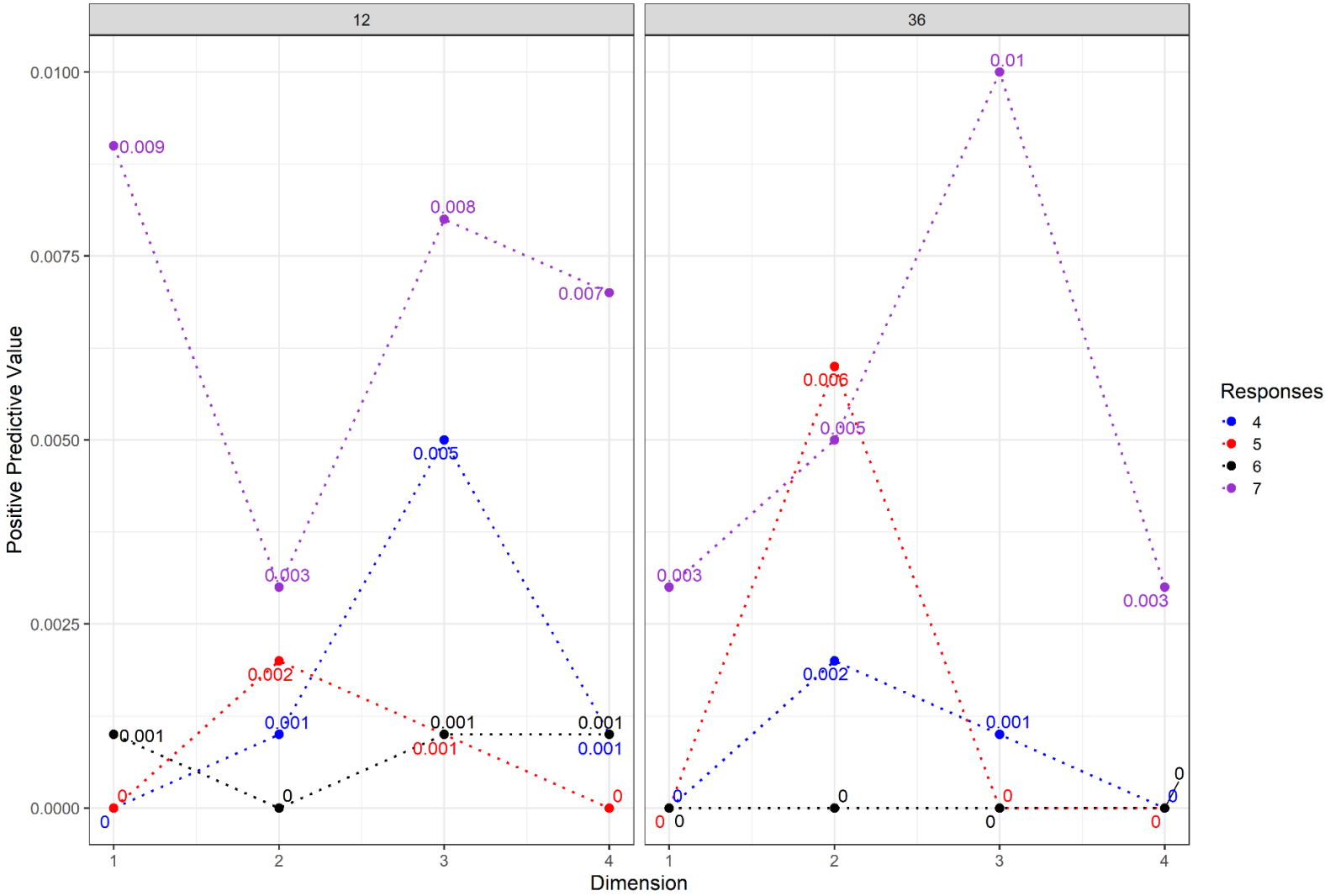


Figure B28

Sensitivity of Guttman Errors by Test Length when Applied to Midpoint Responding

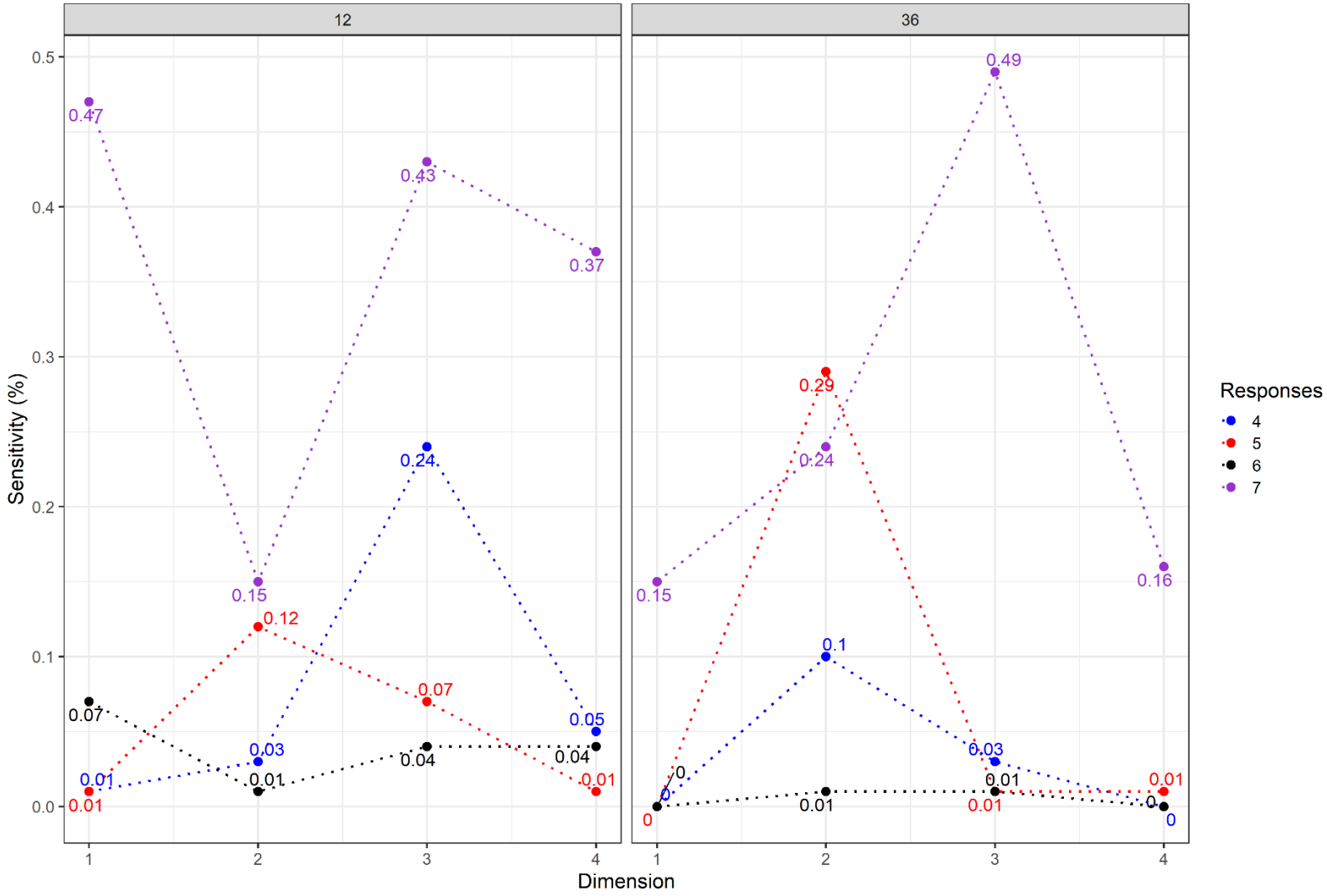


Figure B29

Specificity of Guttman Errors by Test Length when Applied to Midpoint Responding

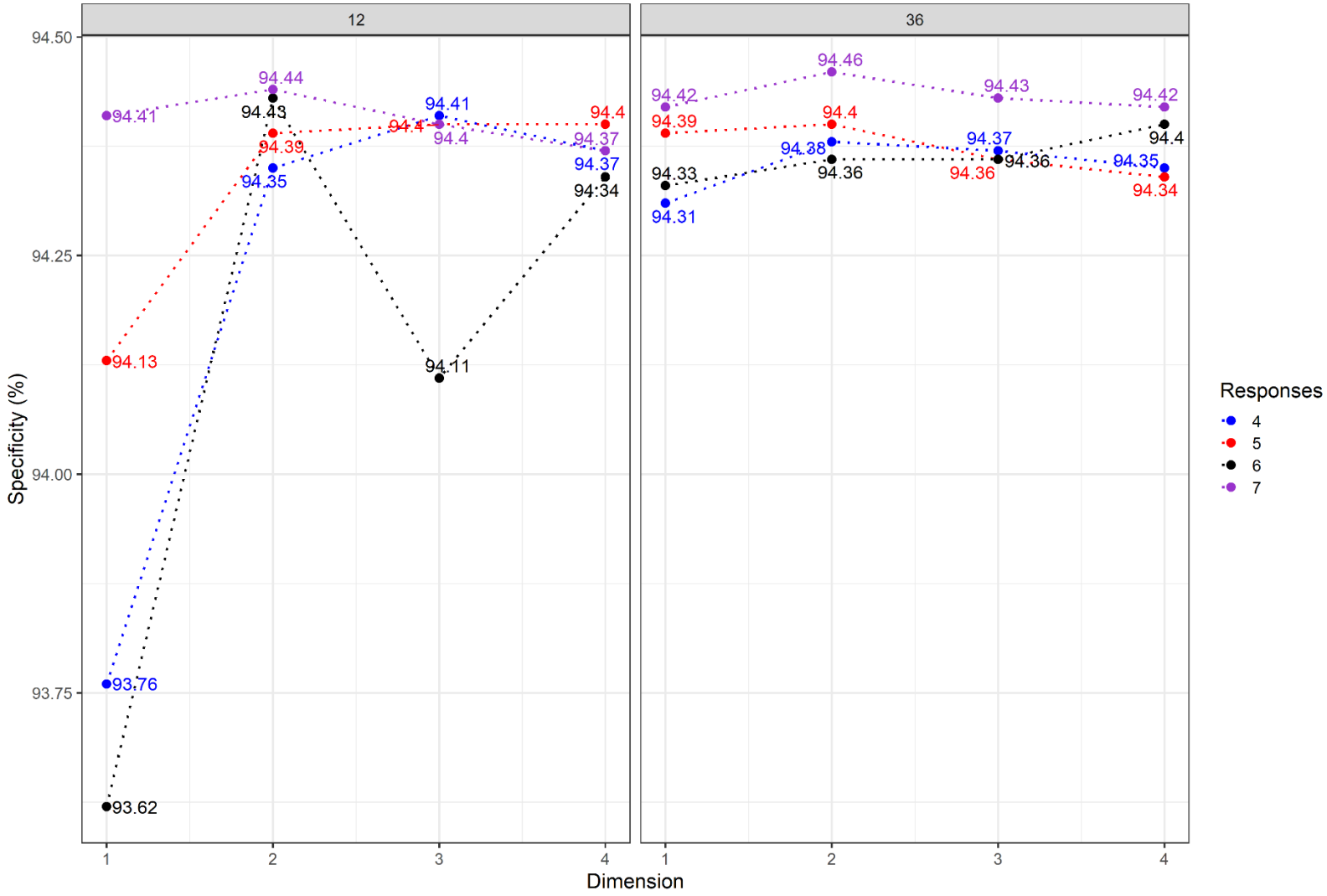


Figure B30

Negative Predictive Value of H^T_i by Test Length when Applied to Midpoint Responding

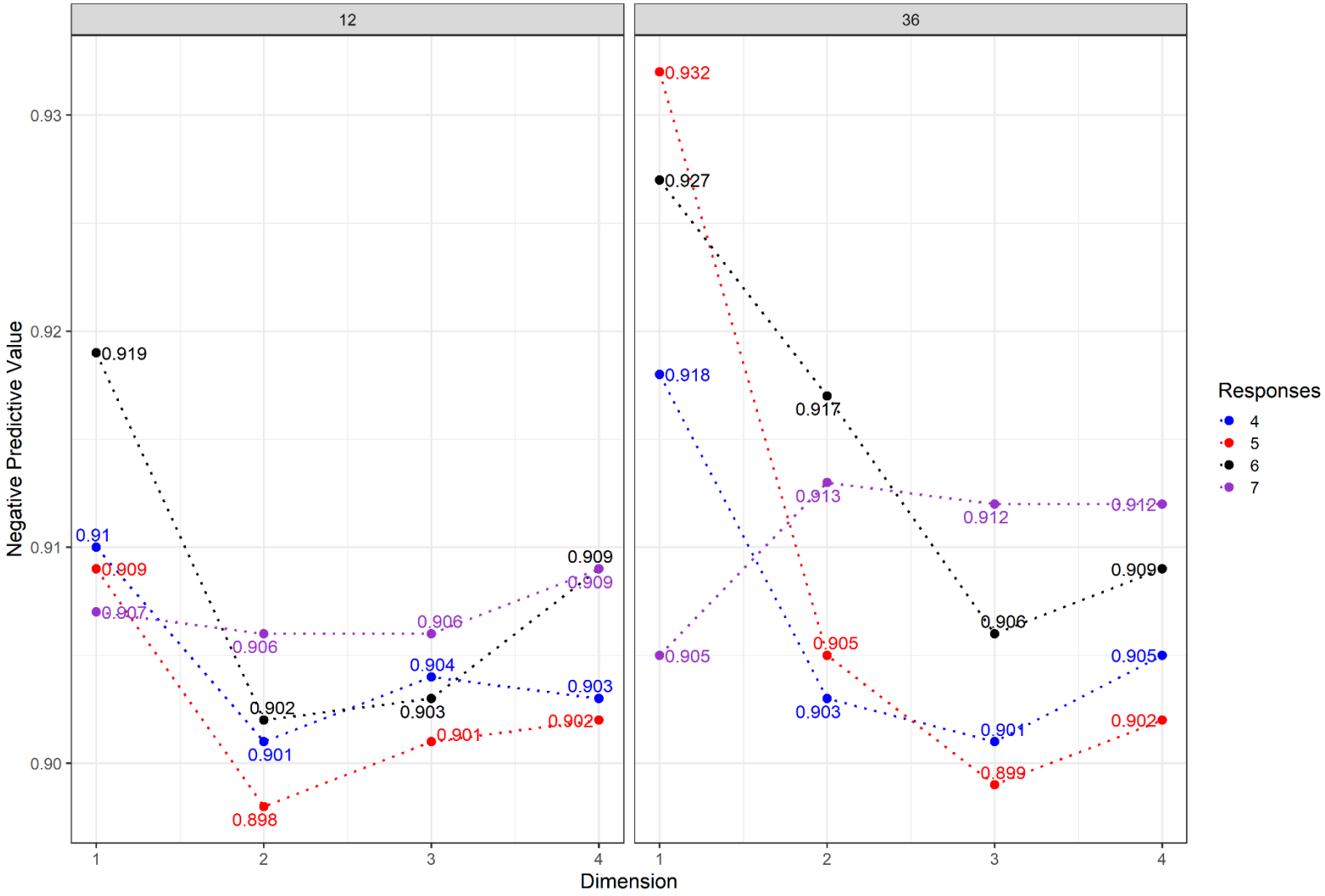


Figure B31

Positive Predictive Value of H^T_i by Test Length when Applied to Midpoint Responding

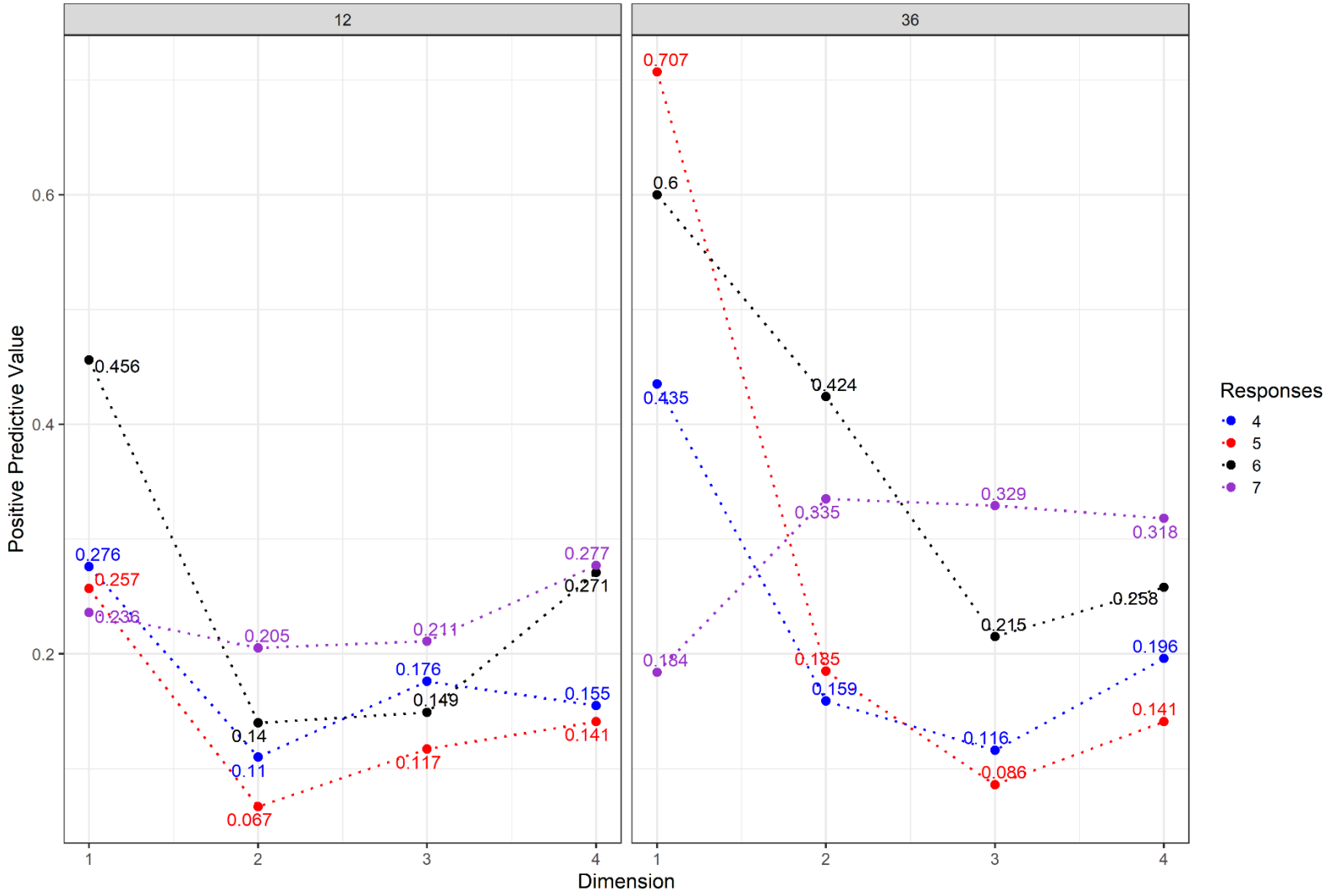


Figure B32

Sensitivity of H^T_i by Test Length when Applied to Midpoint Responding

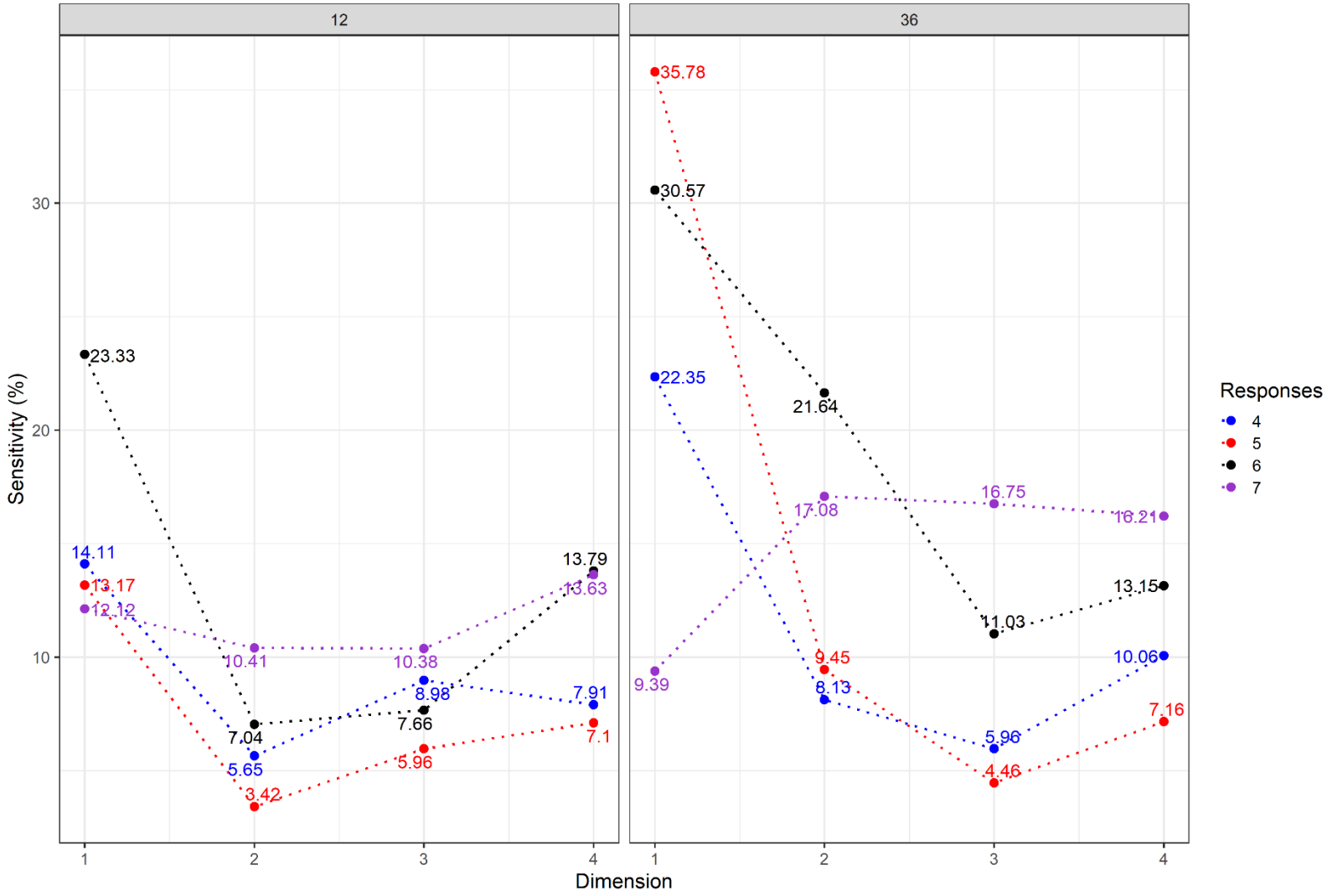


Figure B33

Specificity of H^T_i by Test Length when Applied to Midpoint Responding

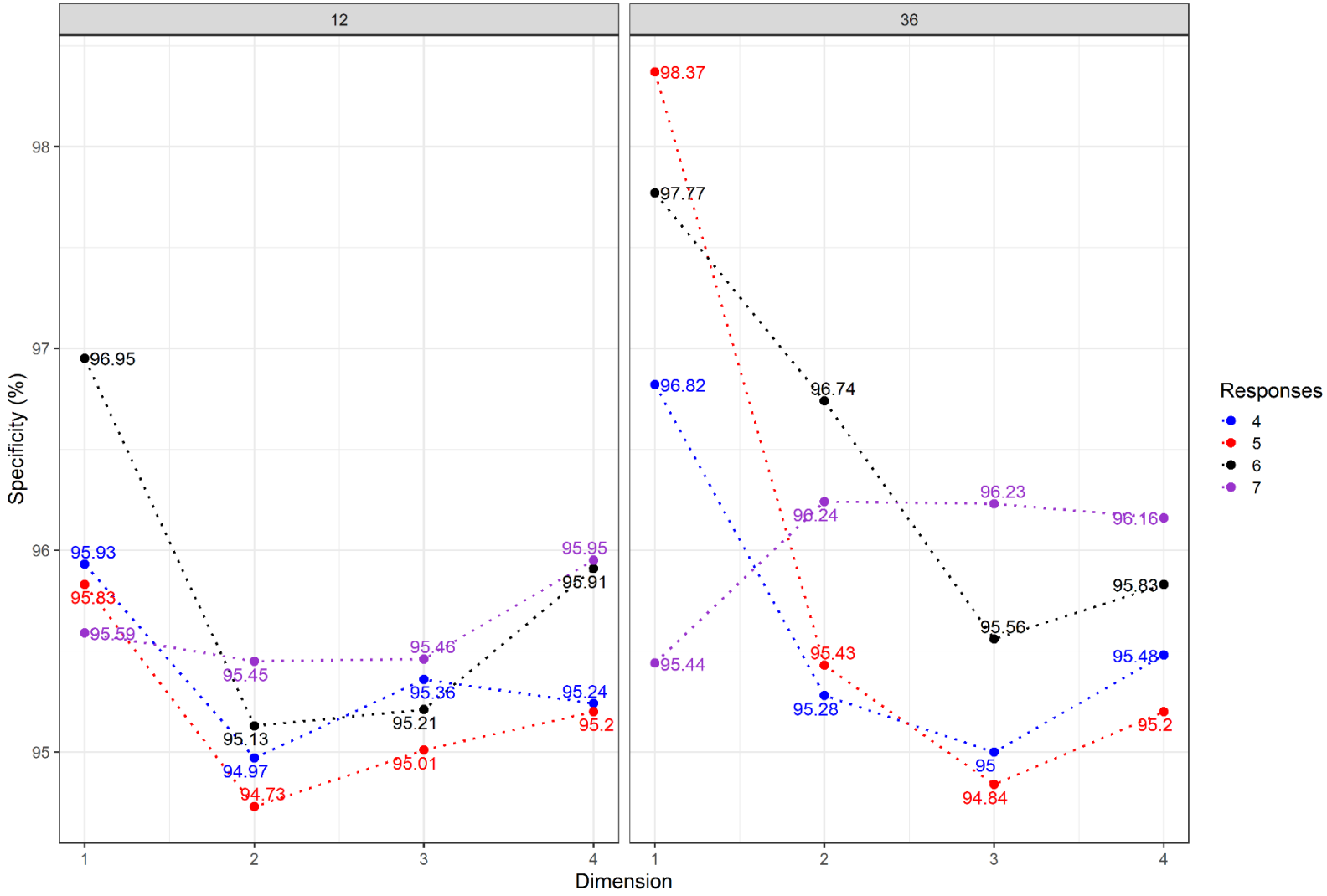


Figure B34

Negative Predictive Value of U3 by Test Length when Applied to Midpoint Responding

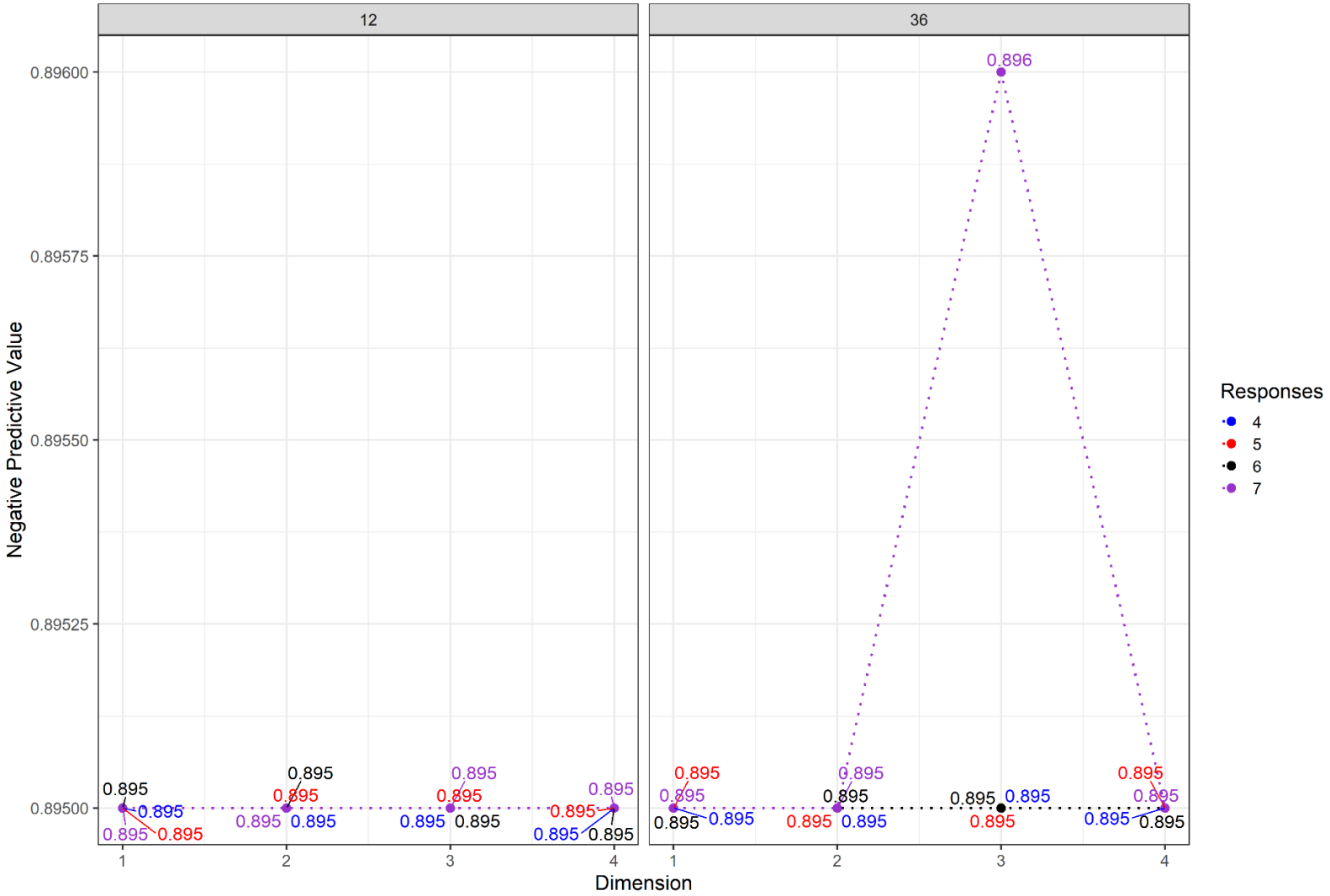


Figure B35

Positive Predictive Value of U3 by Test Length when Applied to Midpoint Responding

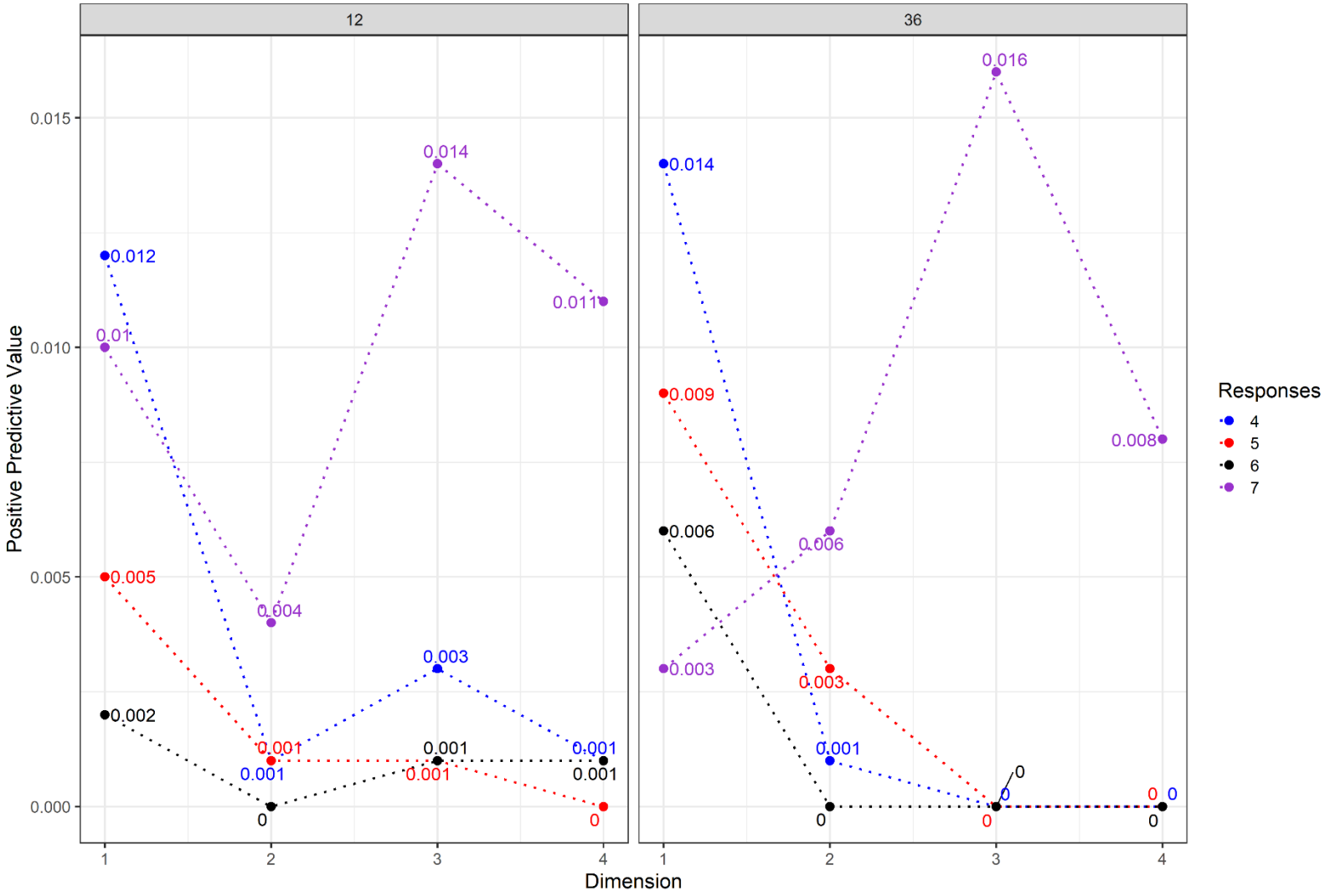


Figure B36

Sensitivity of U3 by Test Length when Applied to Midpoint Responding

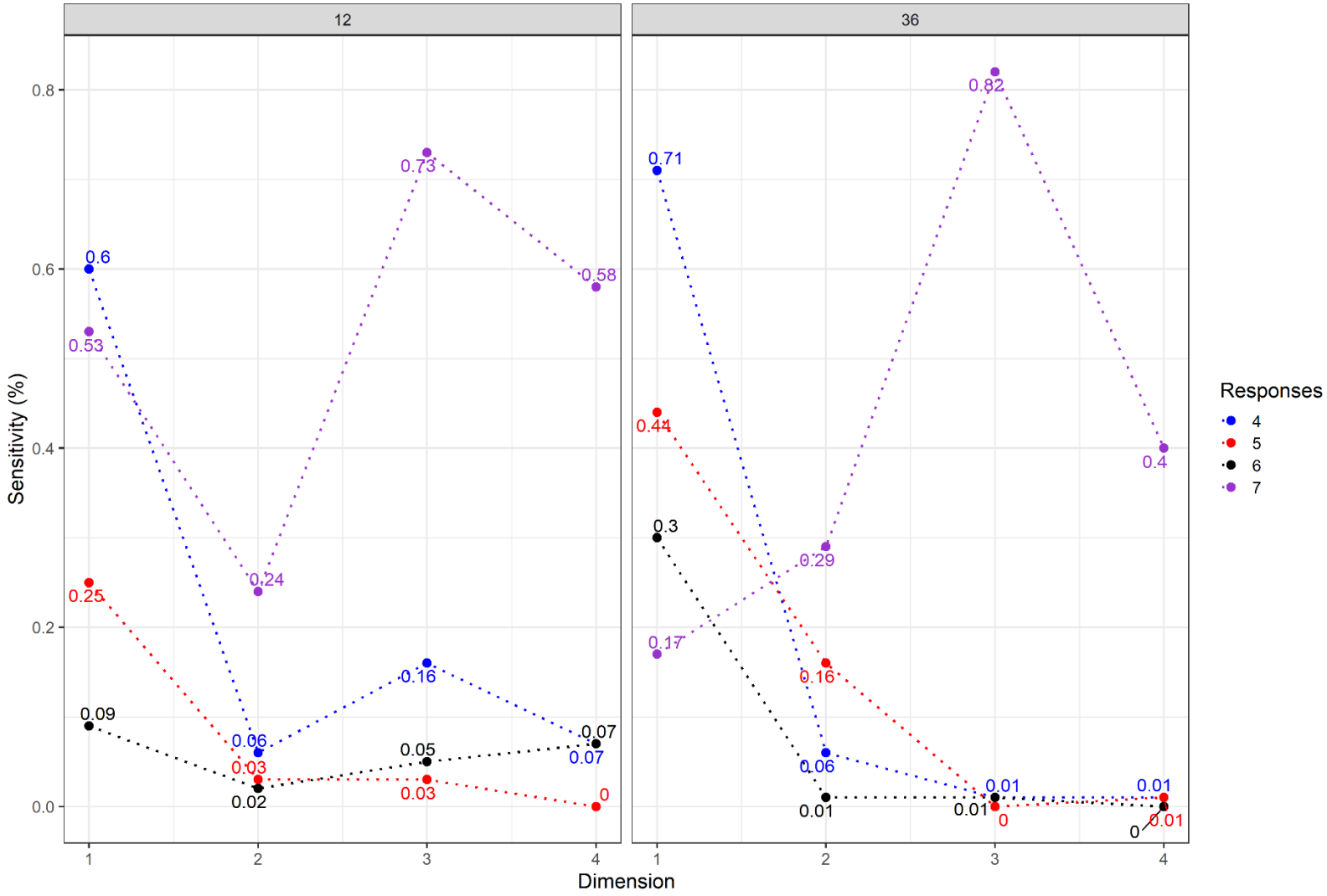


Figure B37

Specificity of U3 by Test Length when Applied to Midpoint Responding

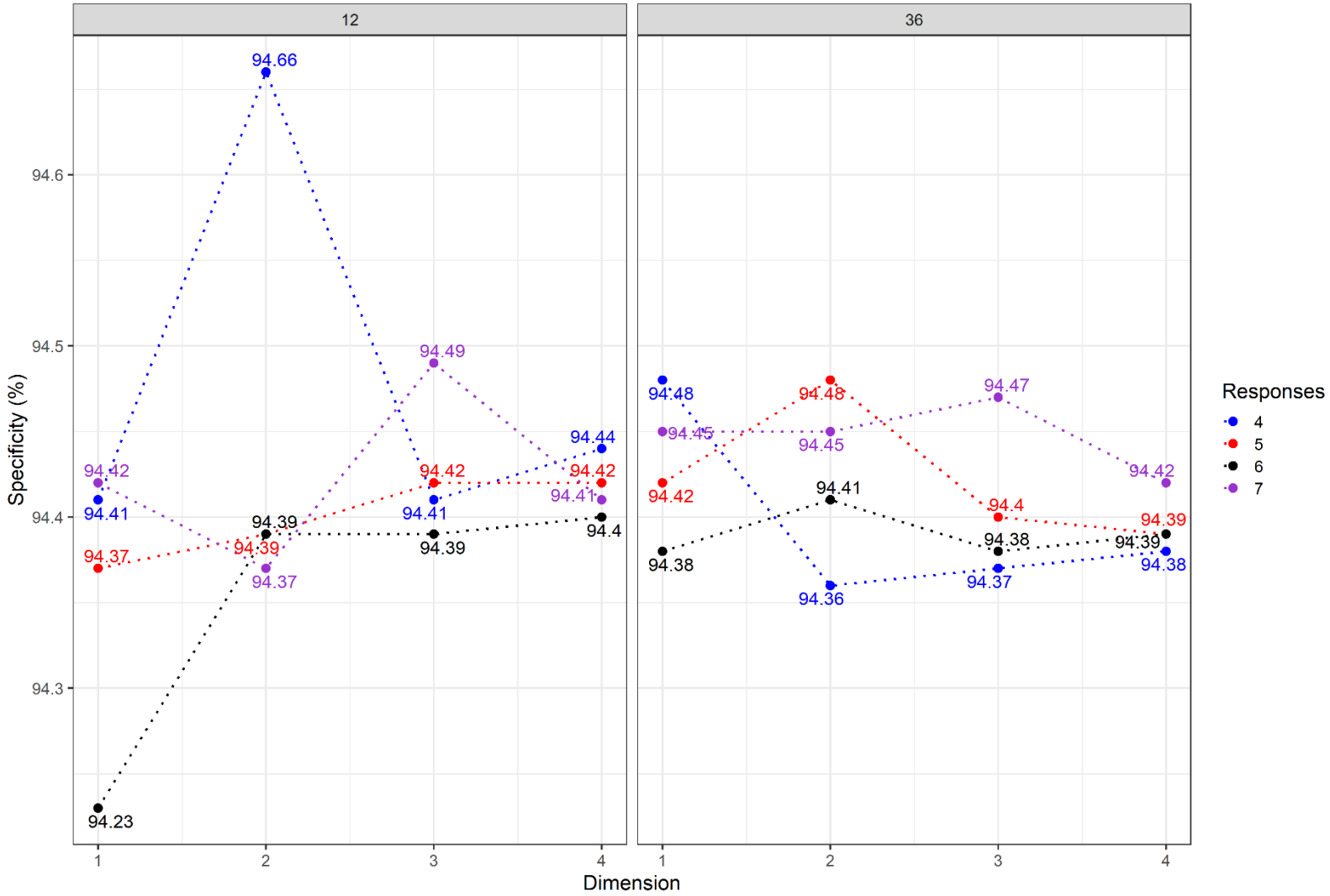


Figure B38

Negative Predictive Value of Guttman Errors by Test Length when Applied to Extreme Responding

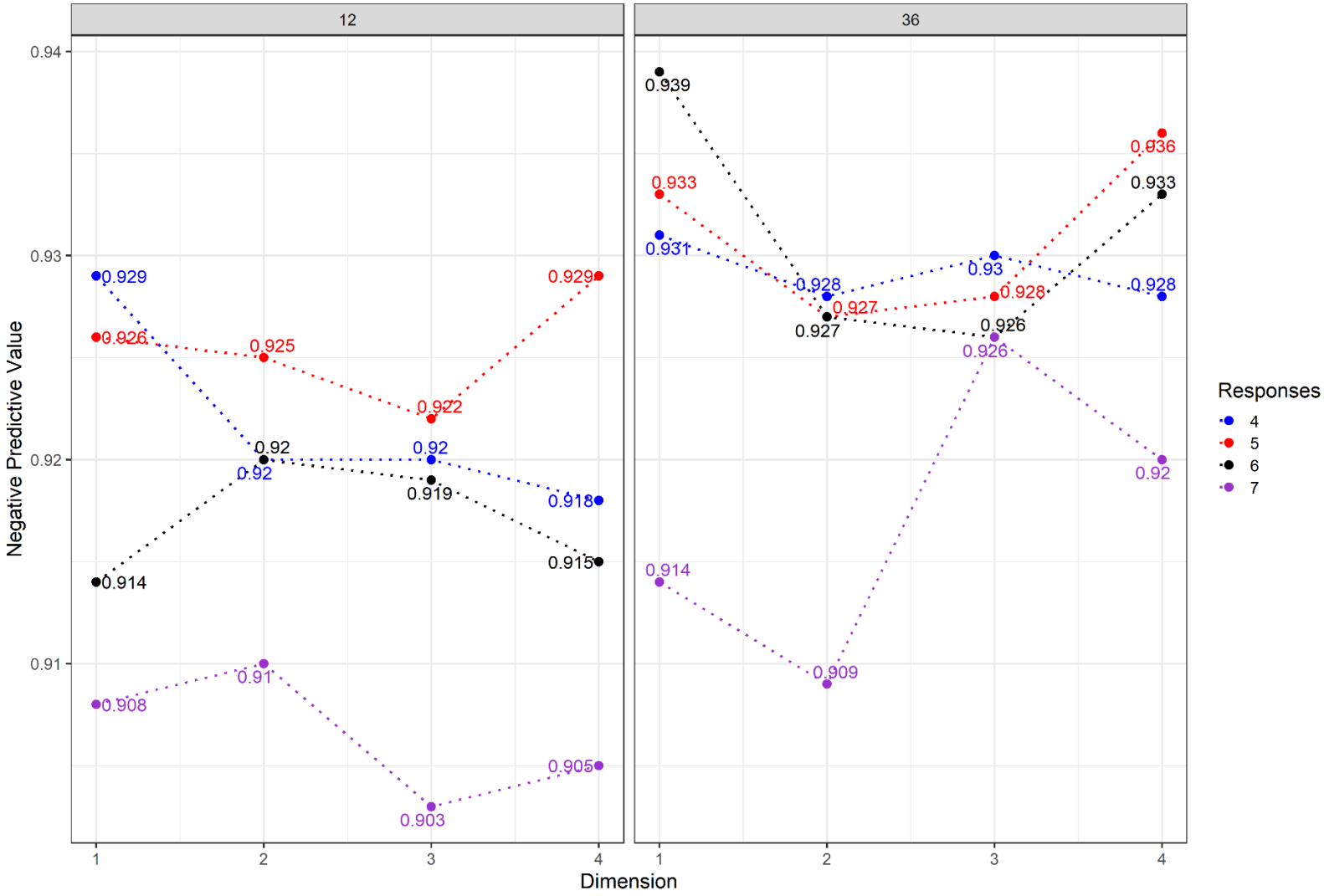


Figure B39

Positive Predictive Value of Guttman Errors by Test Length when Applied to Extreme Responding

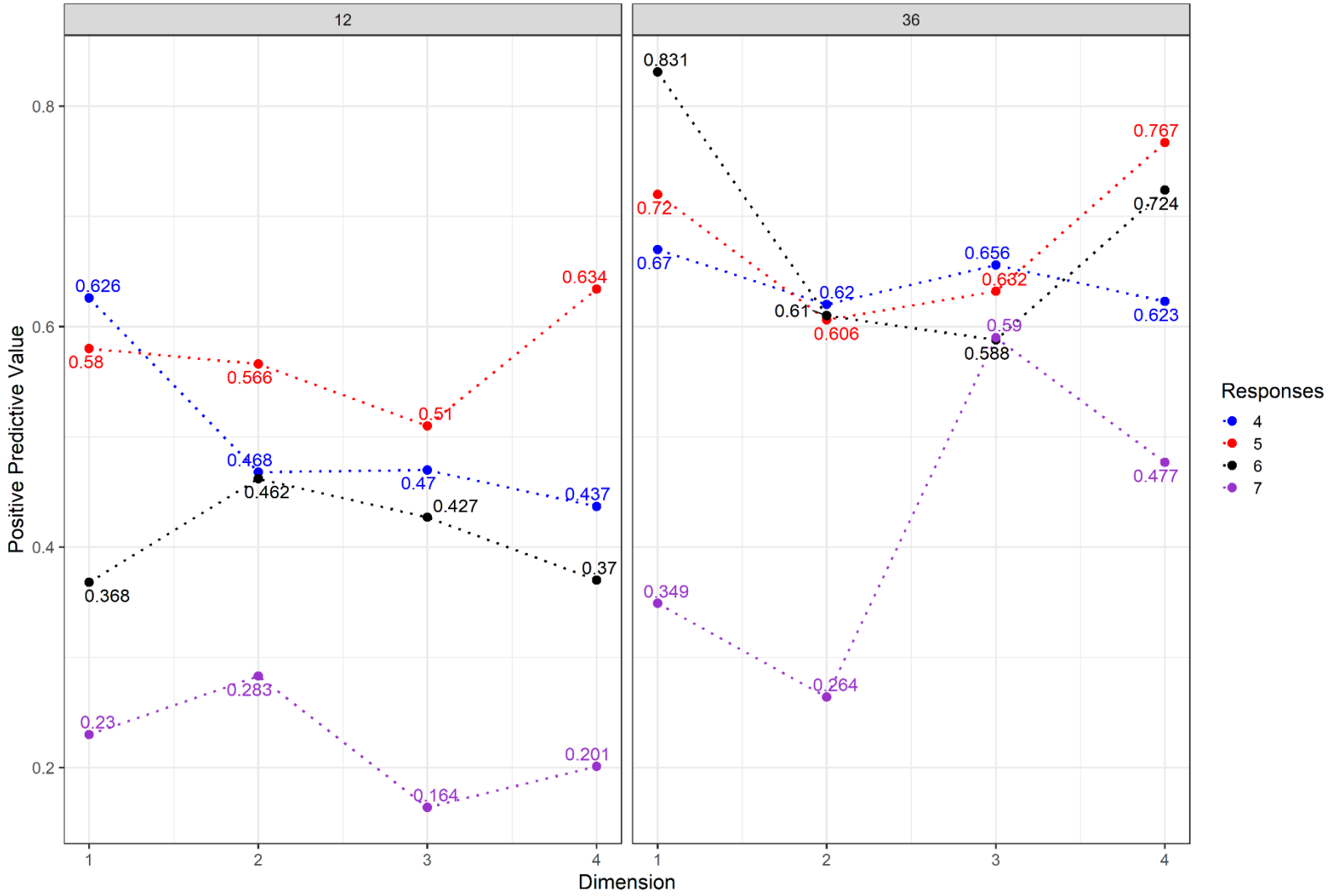


Figure B40

Sensitivity of Guttman Errors by Test Length when Applied to Extreme Responding

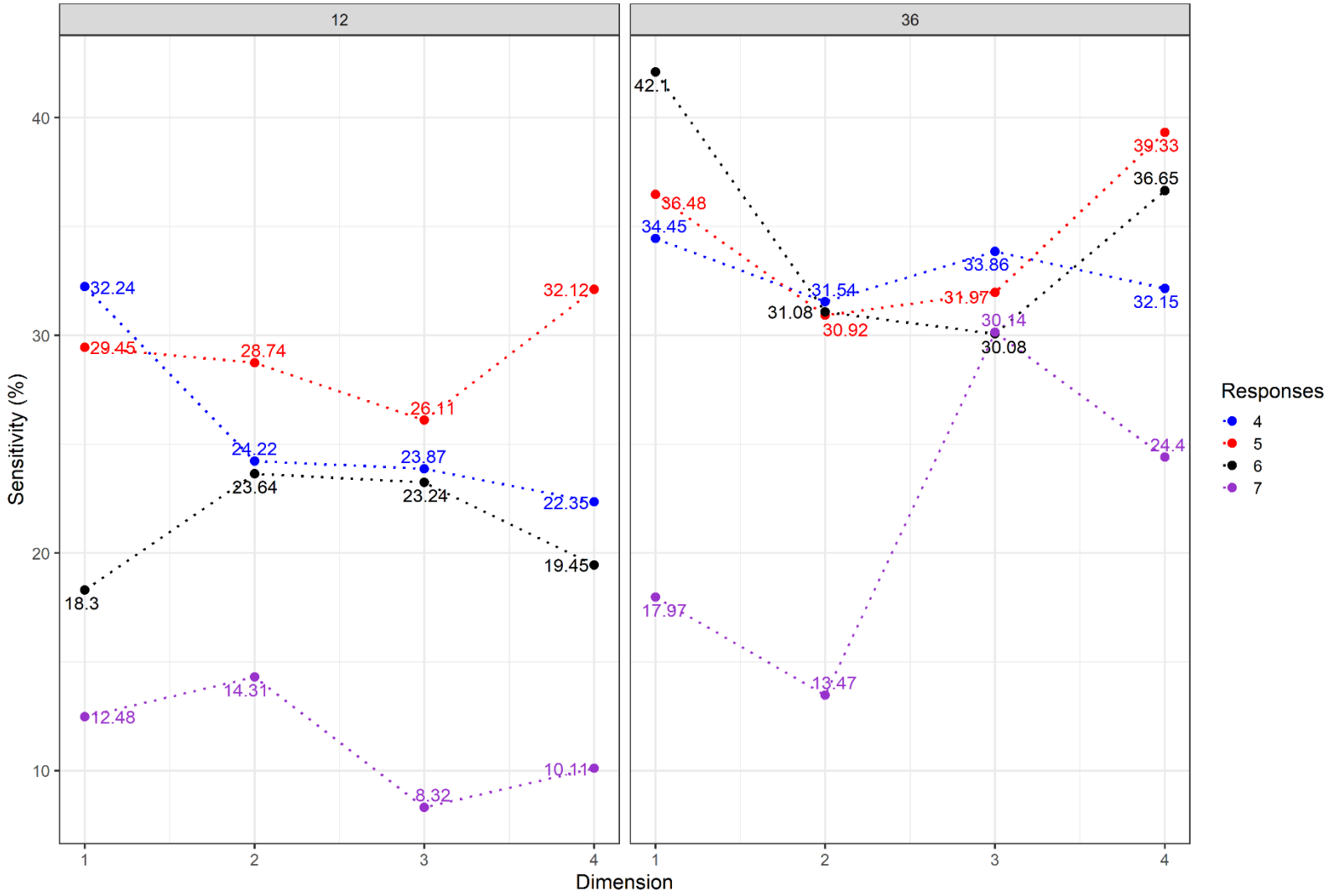


Figure B41

Specificity of Guttman Errors by Test Length when Applied to Extreme Responding

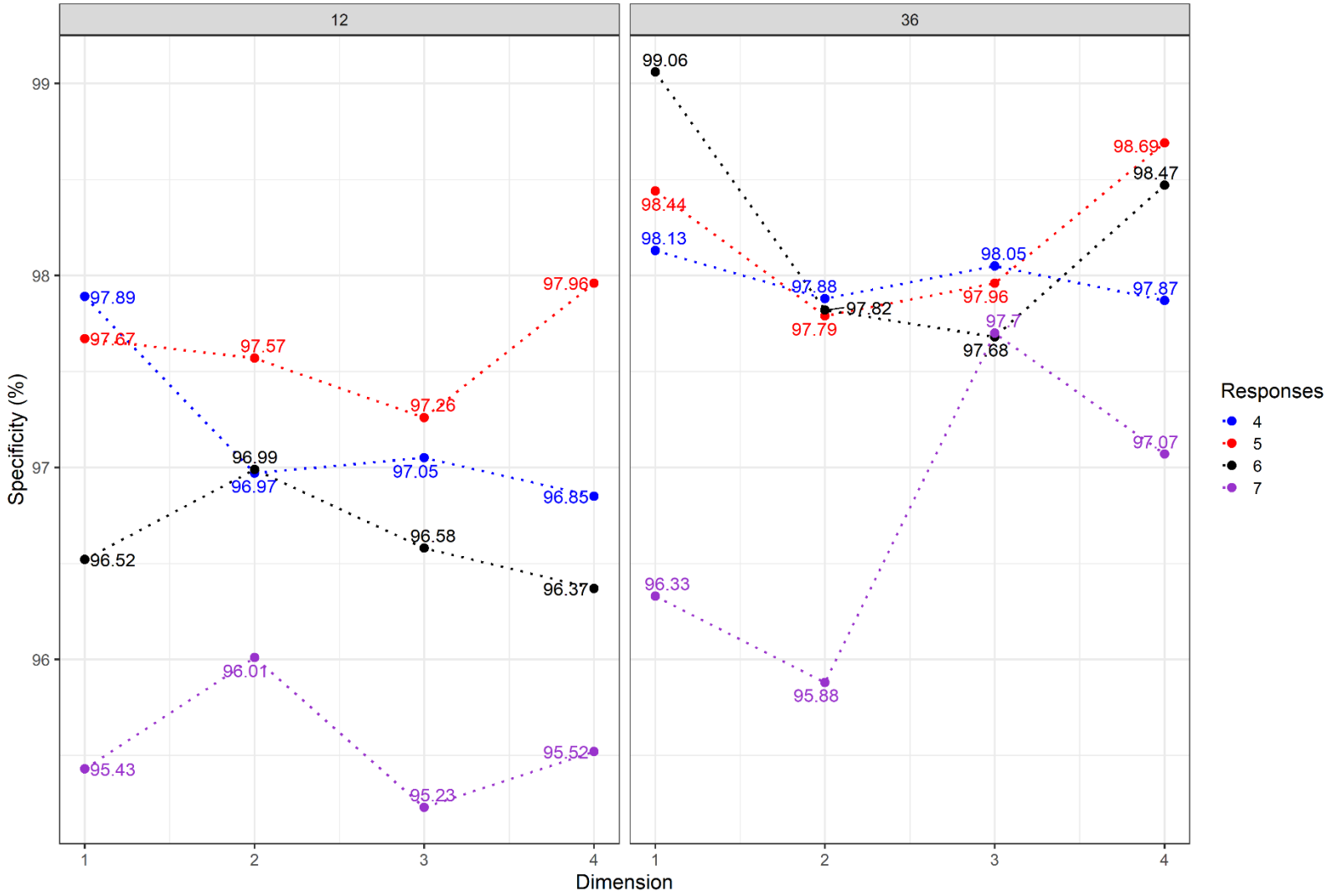


Figure B42

Negative Predictive Value of H^T_i by Test Length when Applied to Extreme Responding

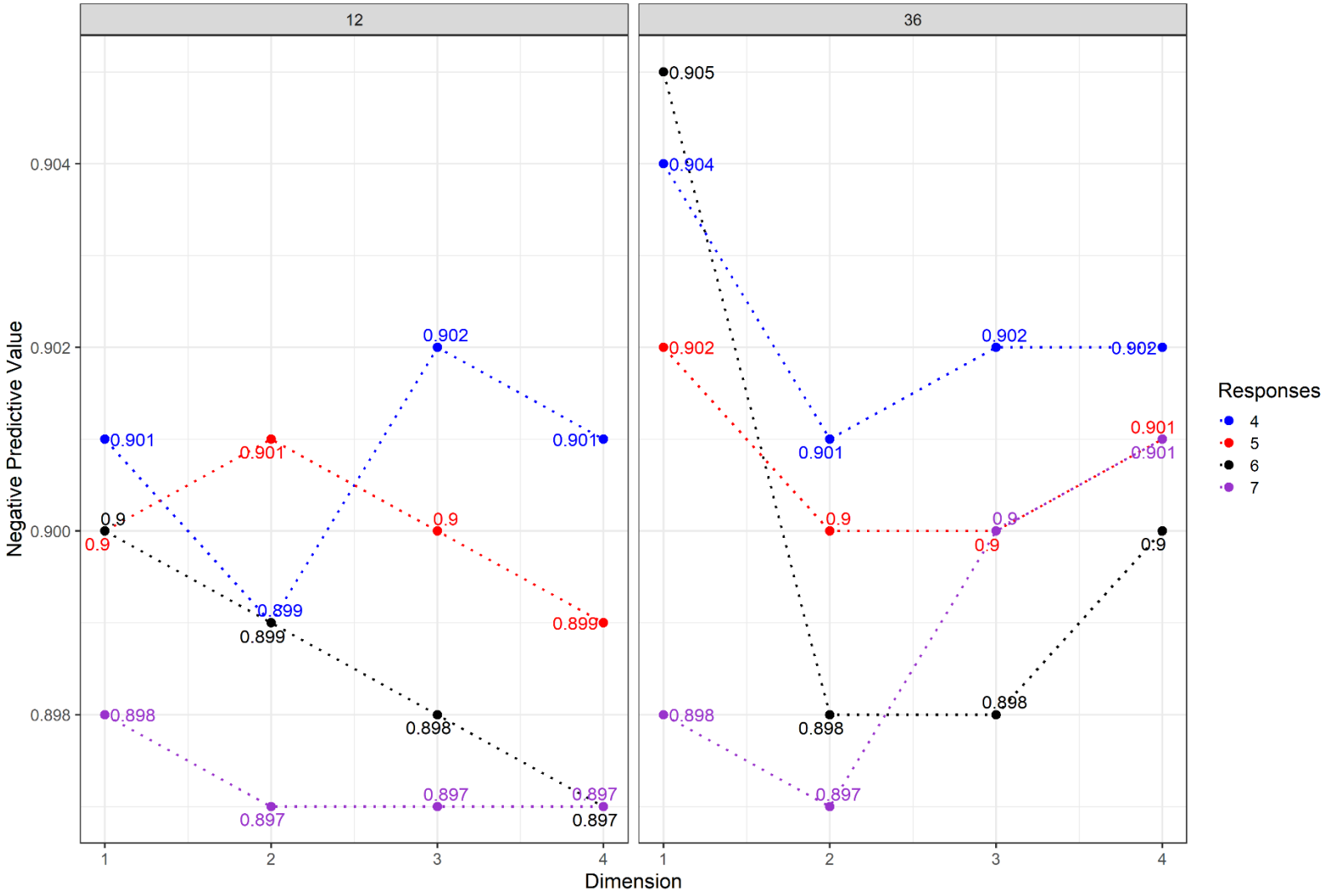


Figure B43

Positive Predictive Value of H^T_i by Test Length when Applied to Extreme Responding

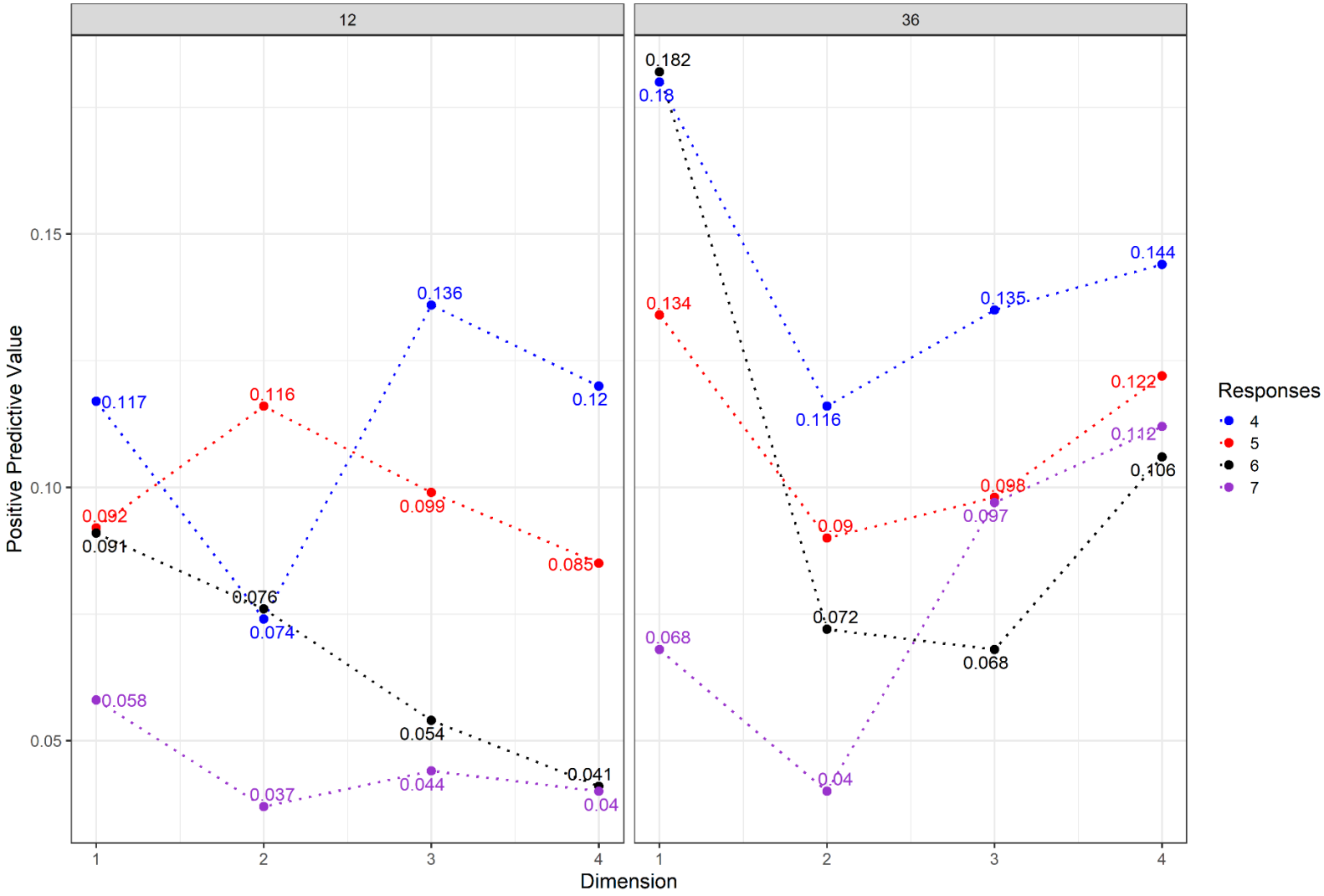


Figure B44

Sensitivity of H^T_i by Test Length when Applied to Extreme Responding

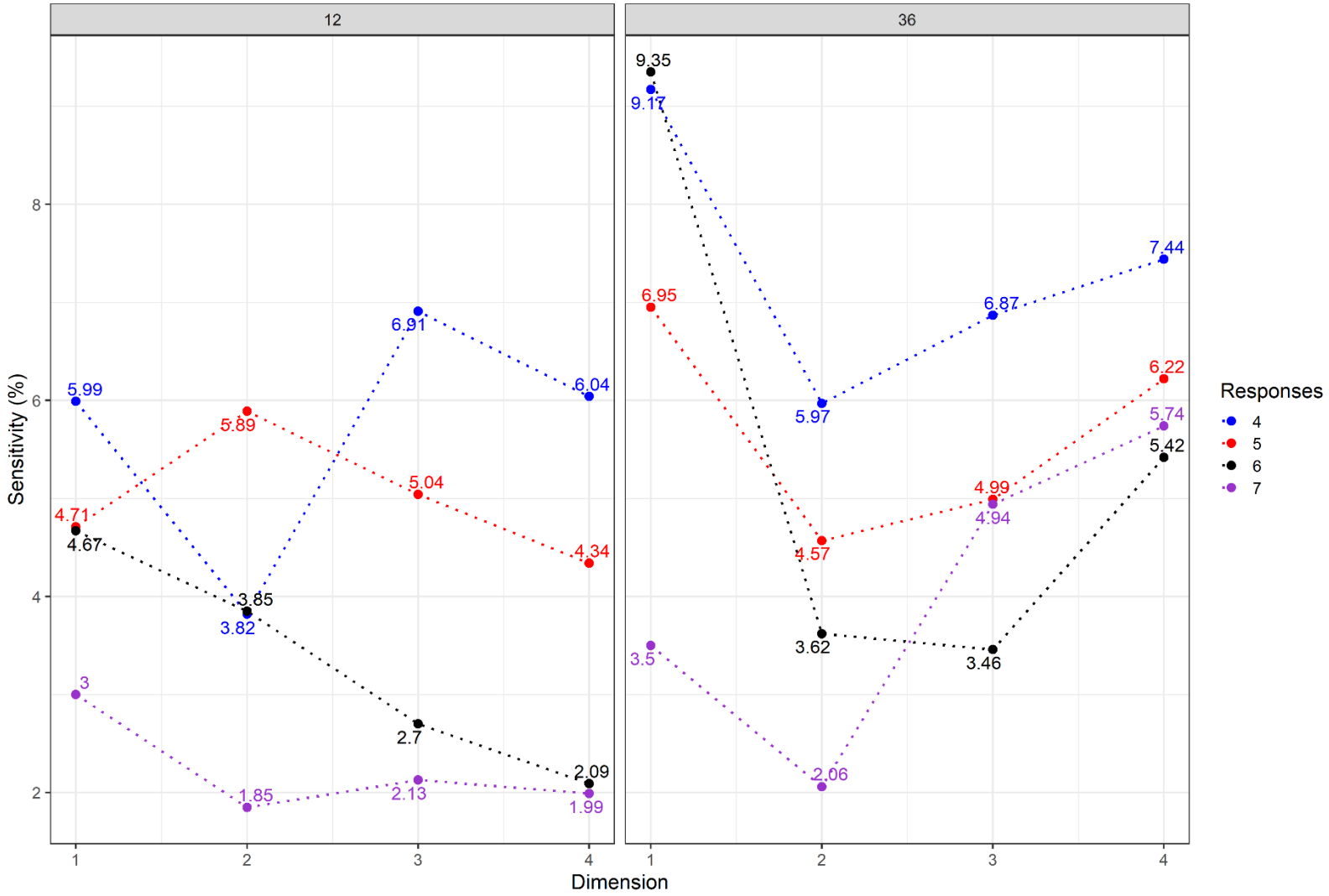


Figure B45

Specificity of H^T_i by Test Length when Applied to Extreme Responding

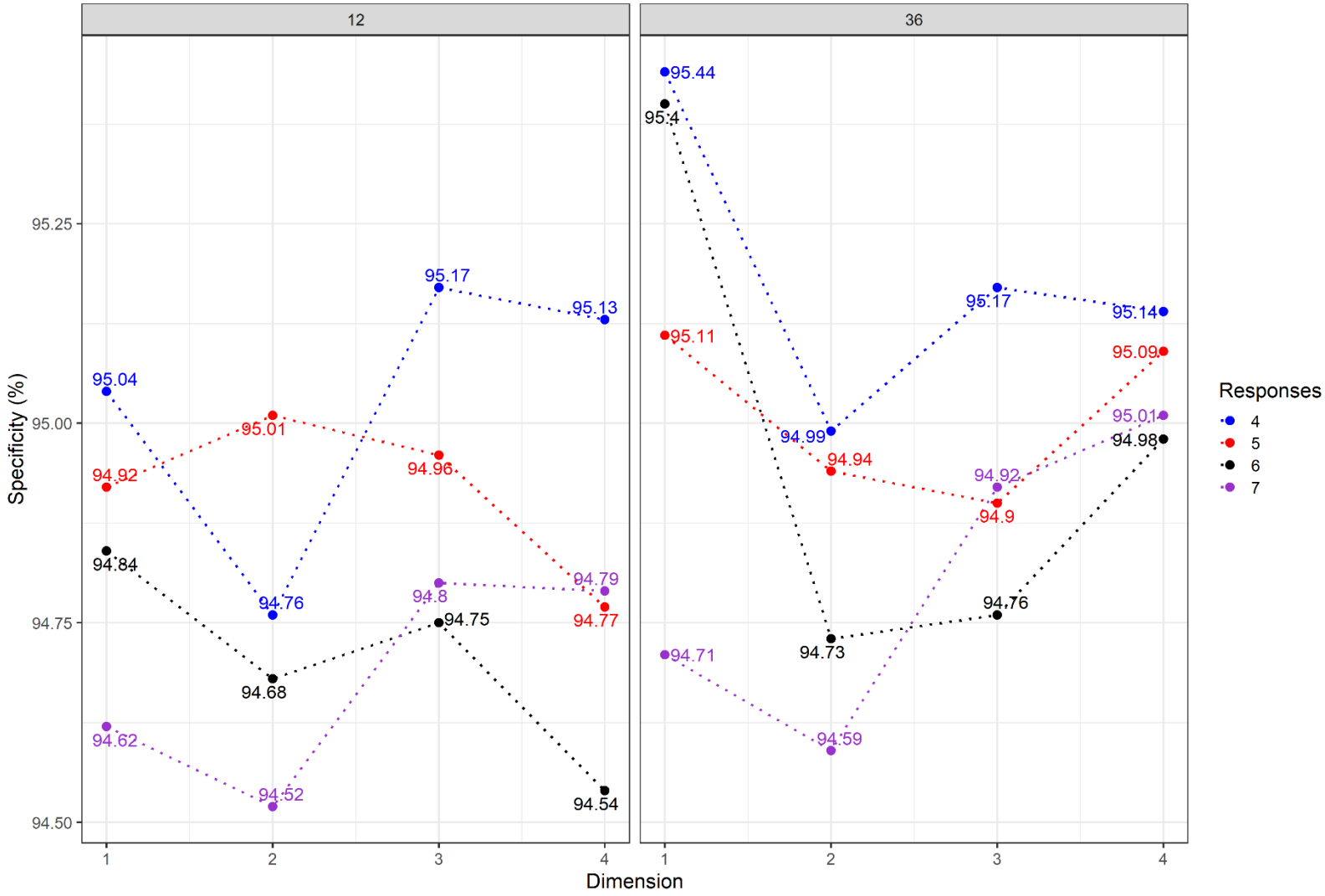


Figure B46

Negative Predictive Value of U3 by Test Length when Applied to Extreme Responding

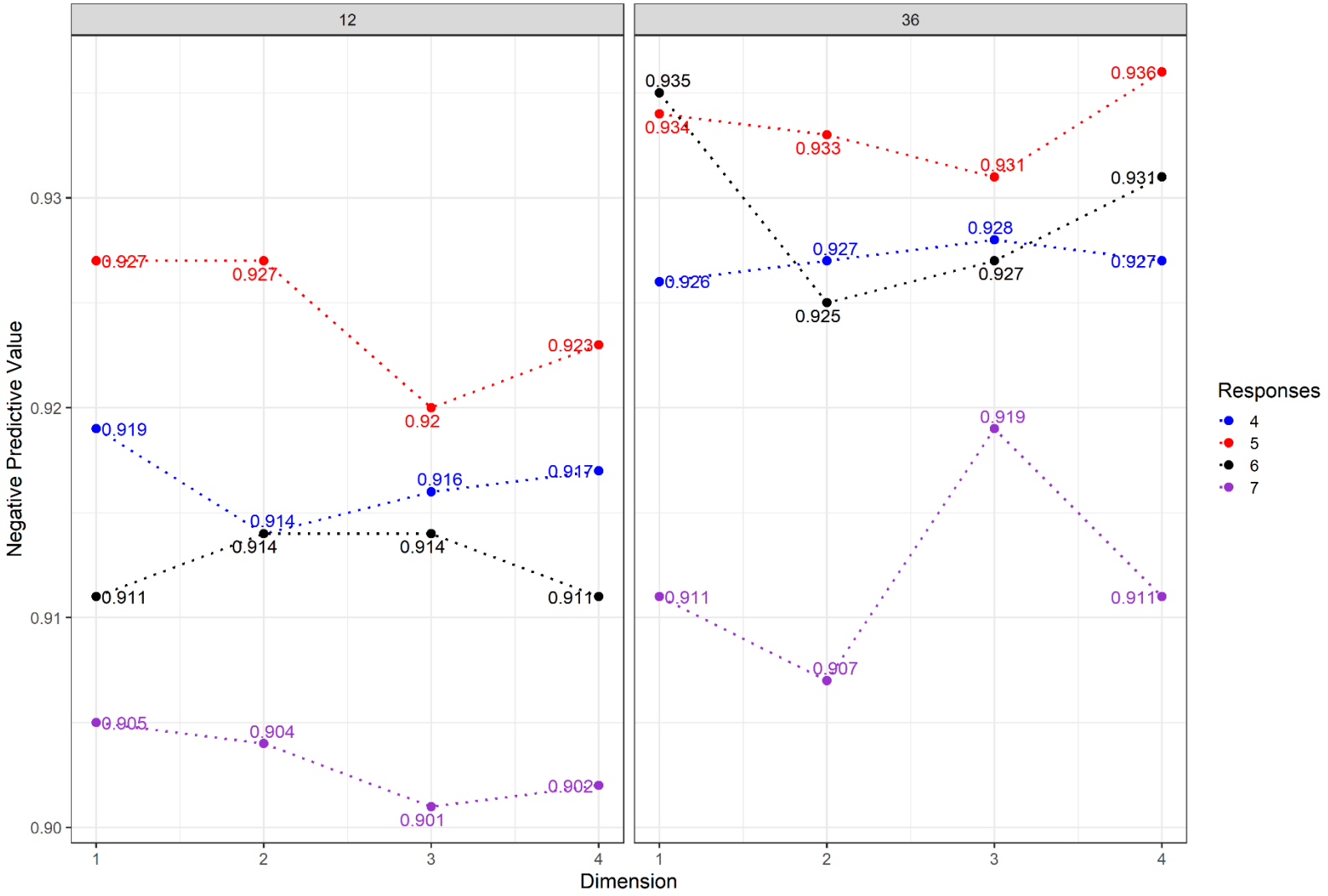


Figure B47

Positive Predictive Value of U3 by Test Length when Applied to Extreme Responding

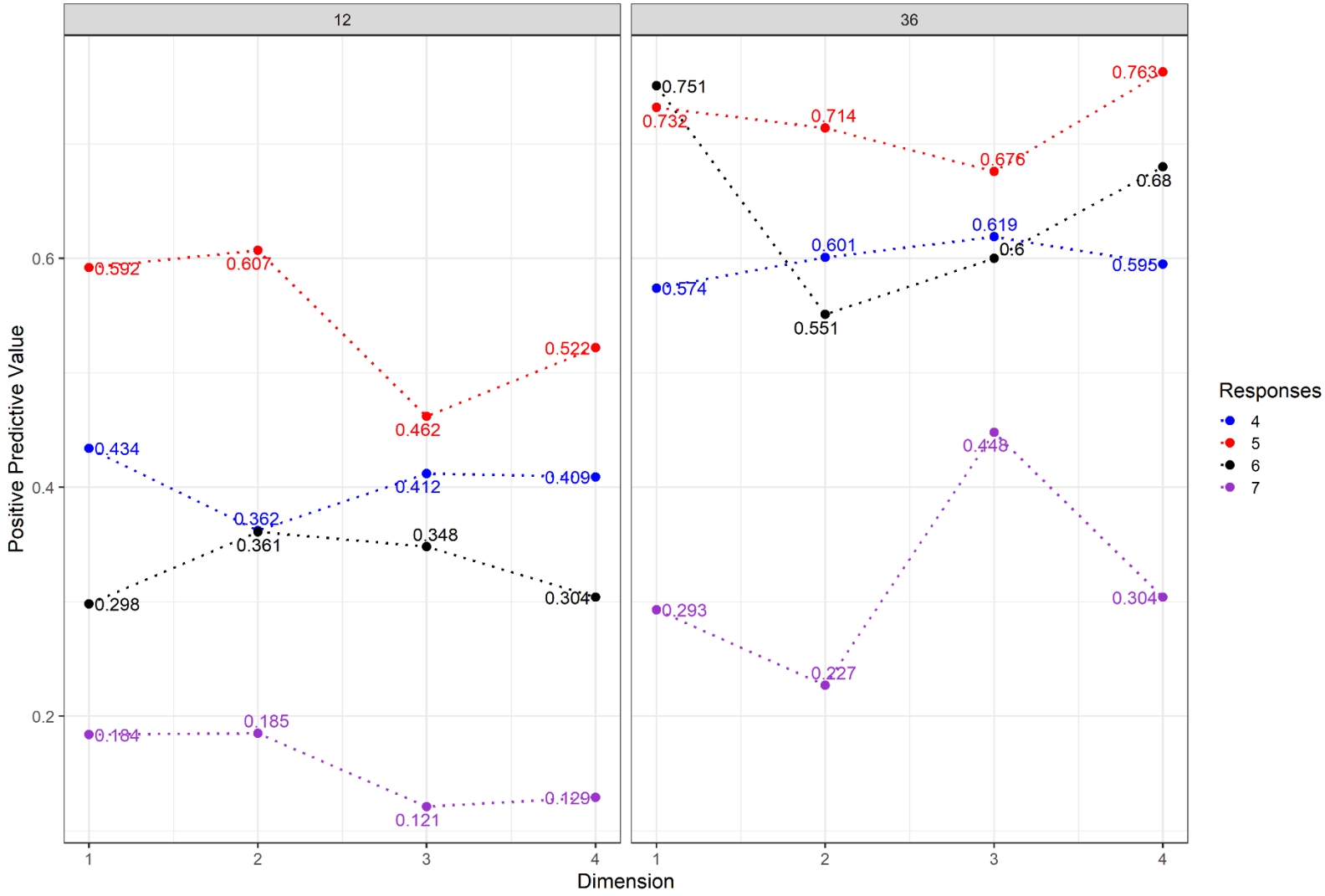


Figure B48

Sensitivity of U3 by Test Length when Applied to Extreme Responding

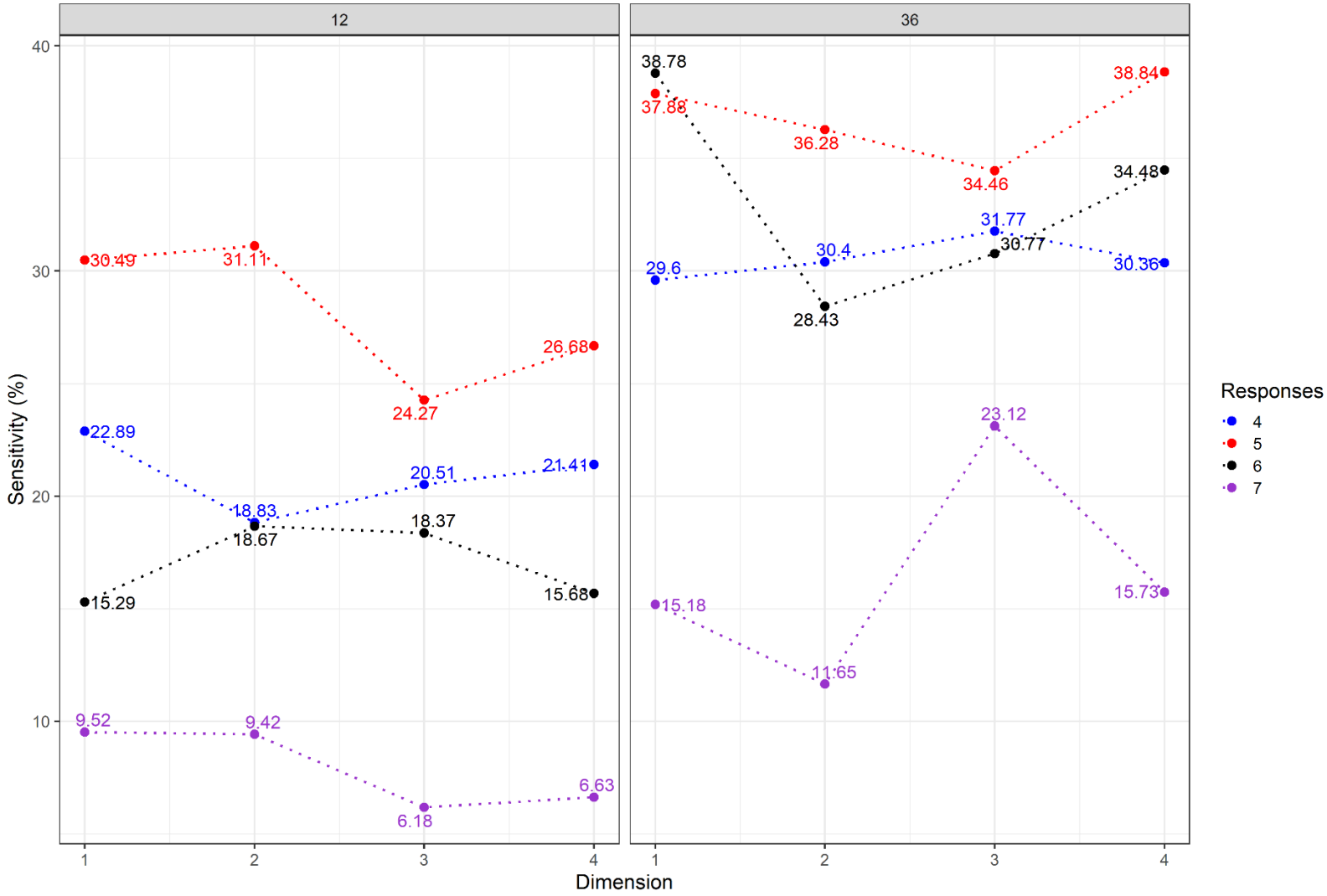


Figure B49

Specificity of U3 by Test Length when Applied to Extreme Responding

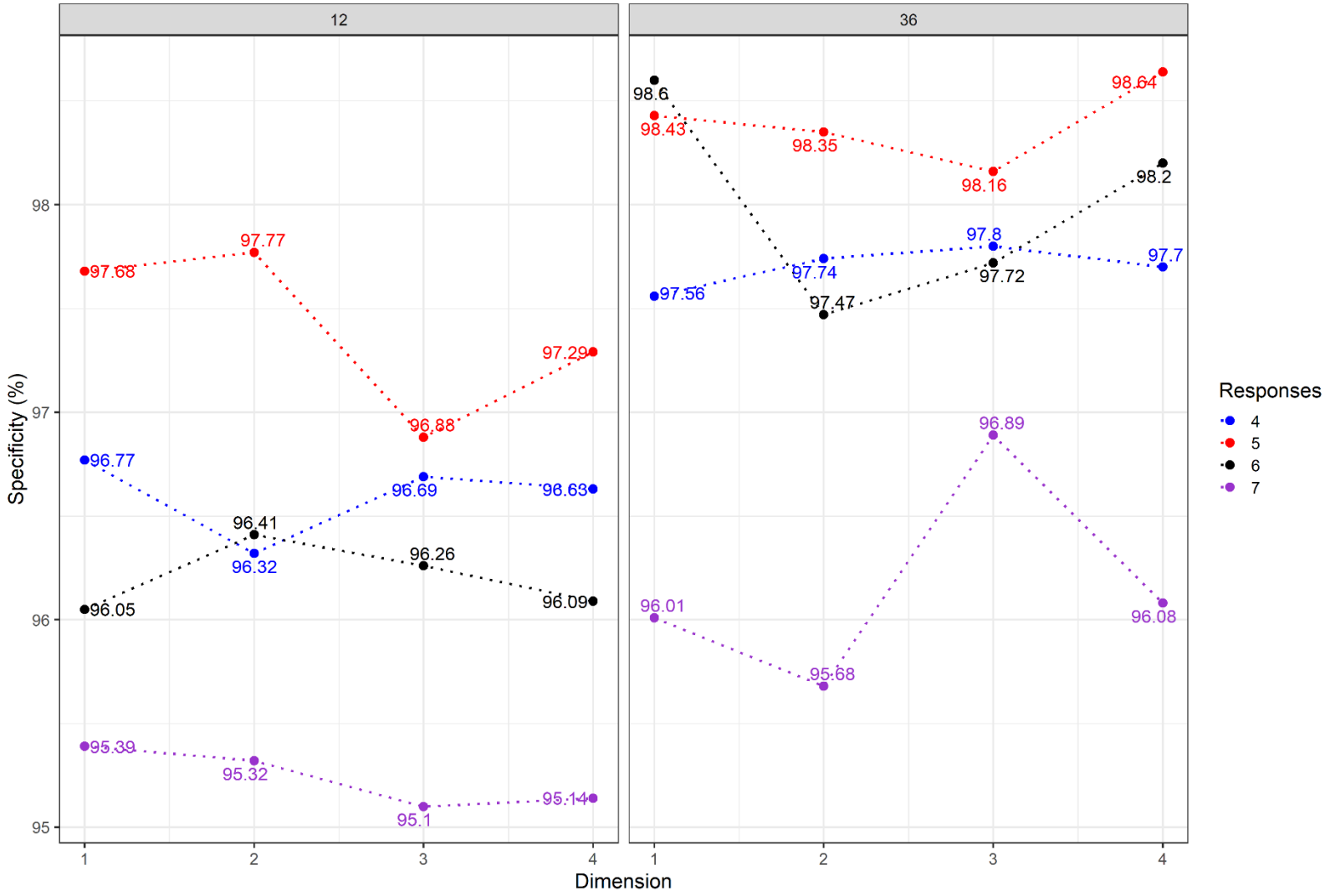


Figure B50

Negative Predictive Value of Guttman Errors by Test Length when Applied to Social Desirability Responding

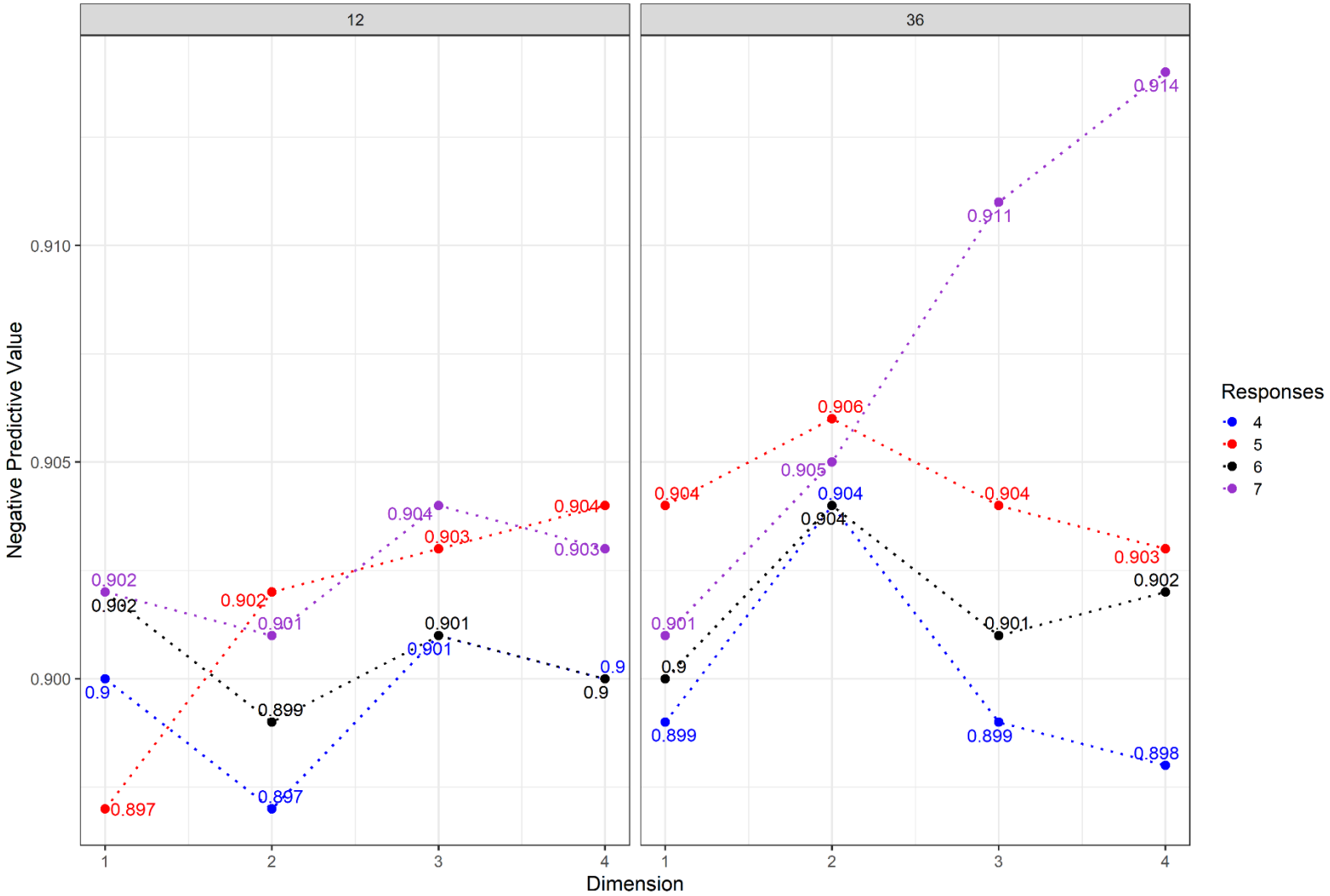


Figure B51

Positive Predictive Value of Guttman Errors by Test Length when Applied to Social Desirability Responding

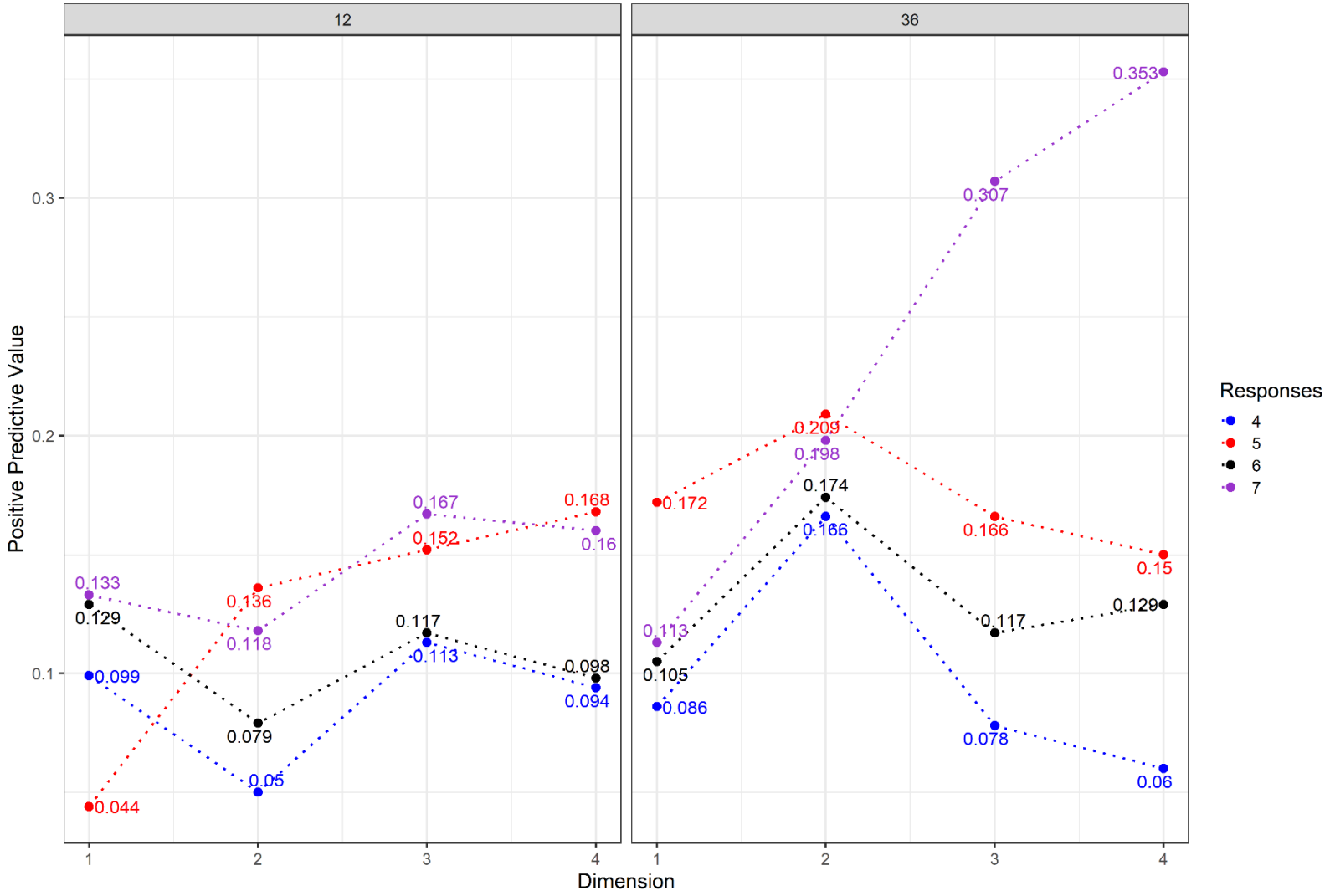


Figure B52

Sensitivity of Guttman Errors by Test Length when Applied to Social Desirability Responding

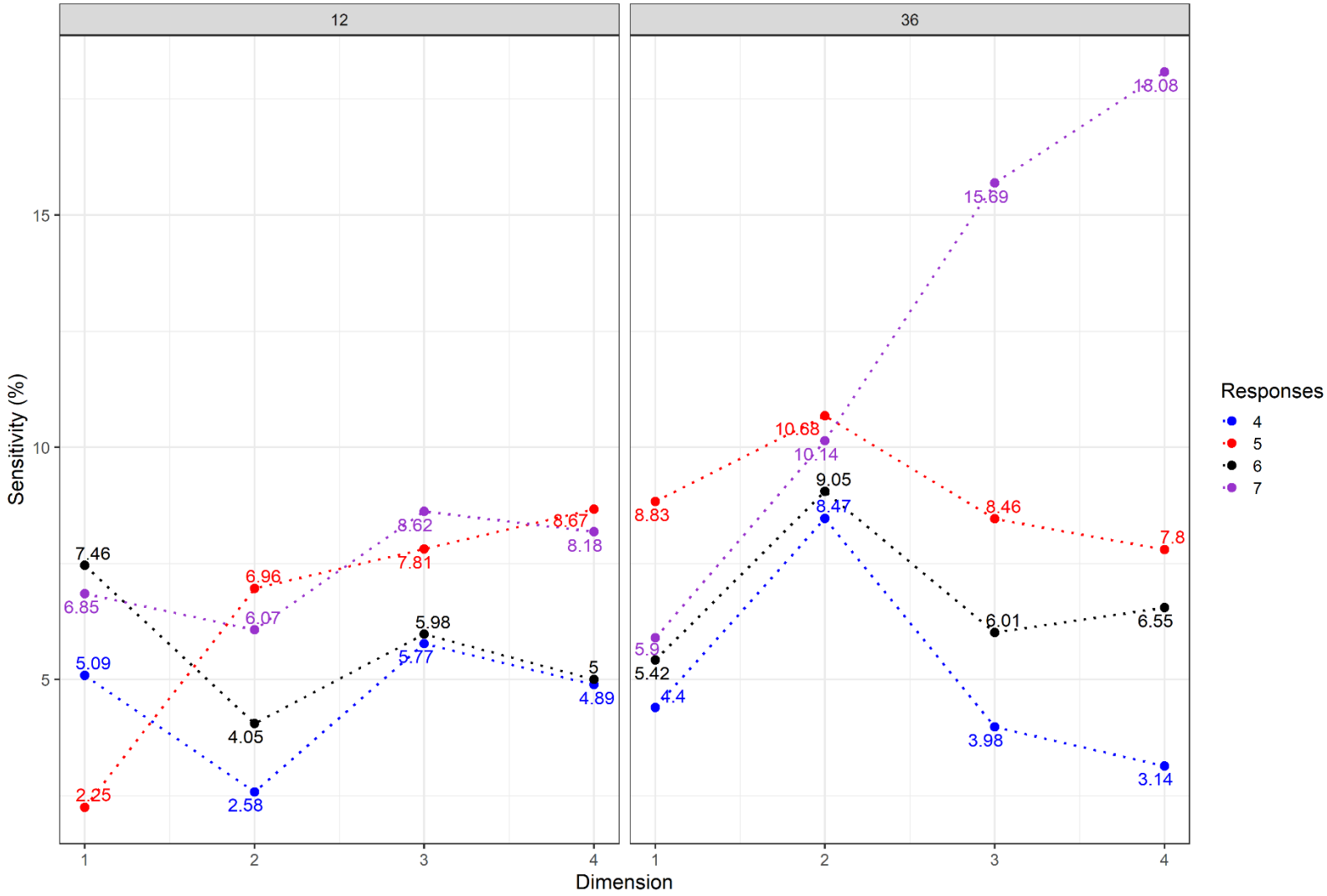


Figure B53

Specificity of Guttman Errors by Test Length when Applied to Social Desirability Responding

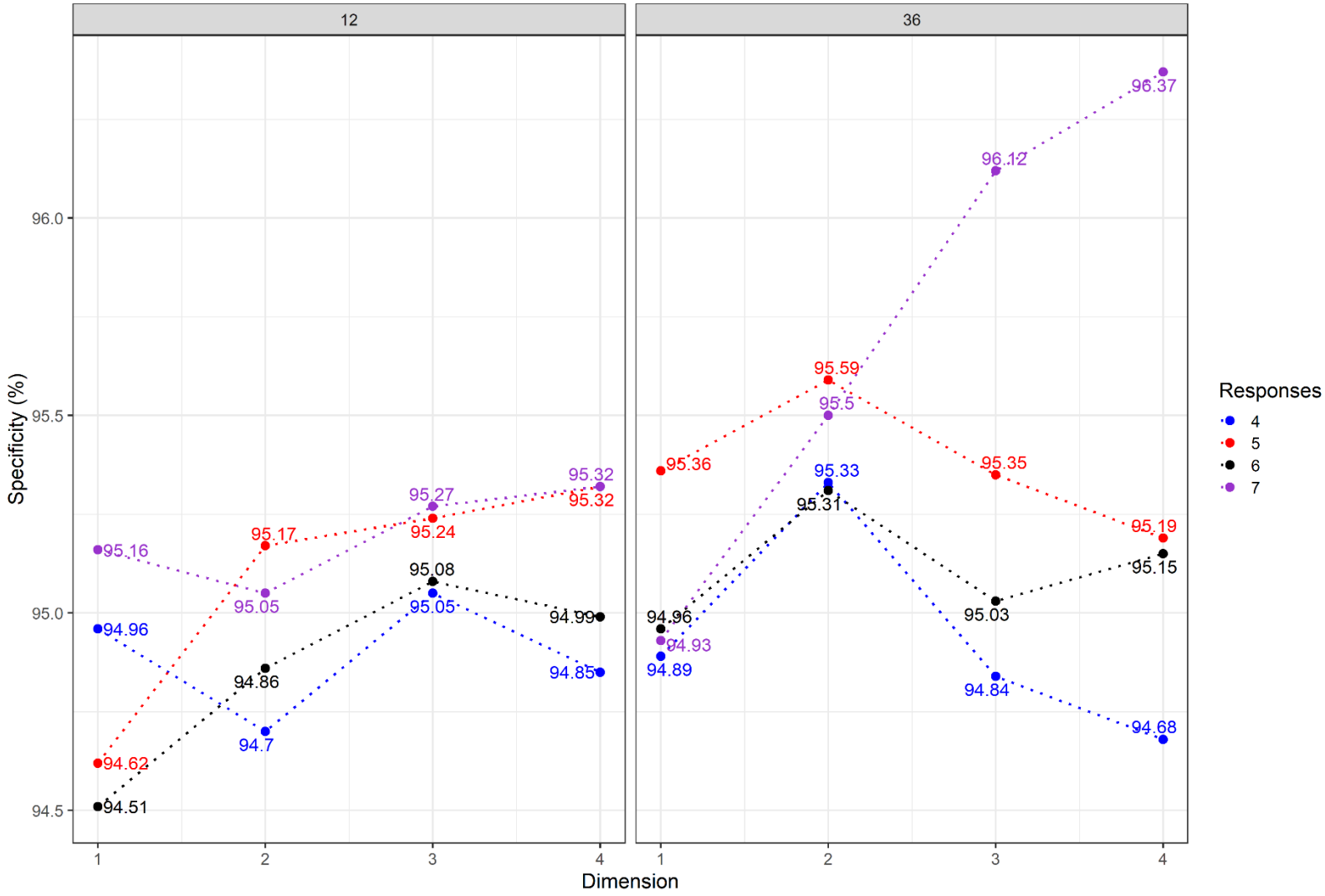


Figure B54

Negative Predictive Value of H^T_i by Test Length when Applied to Social Desirability Responding

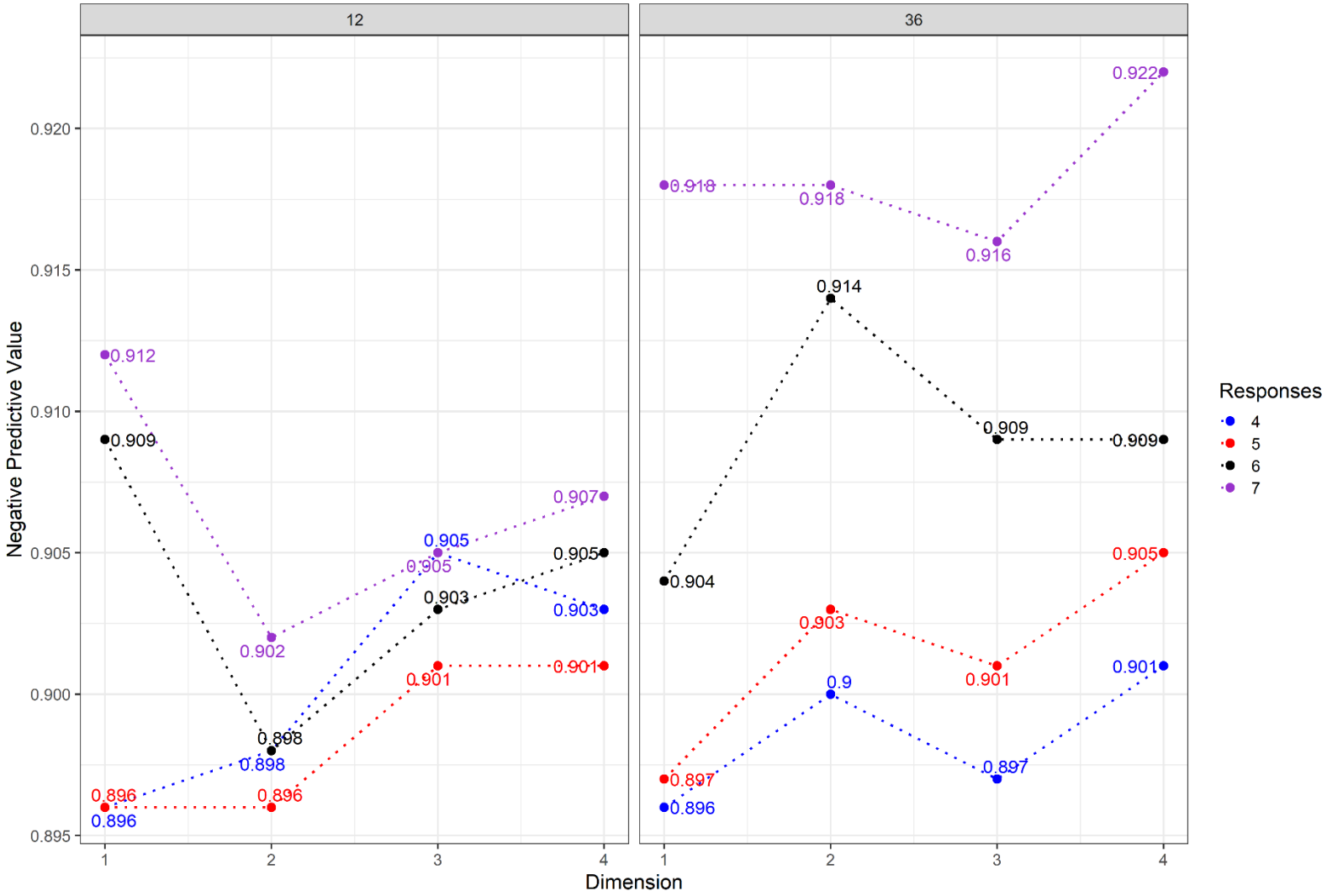


Figure B55

Positive Predictive Value of H^T_i by Test Length when Applied to Social Desirability Responding

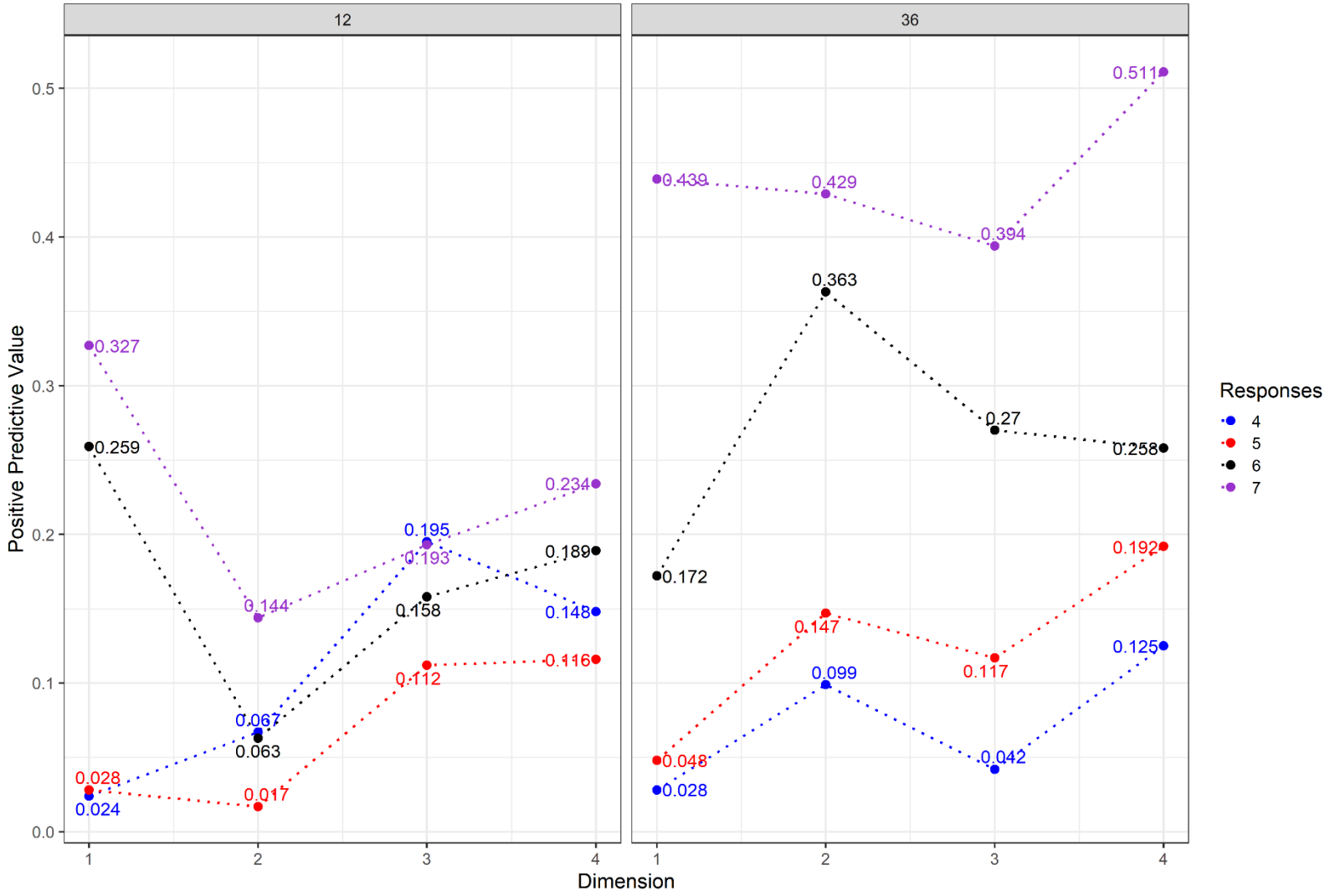


Figure B56

Sensitivity of H^T_i by Test Length when Applied to Social Desirability Responding

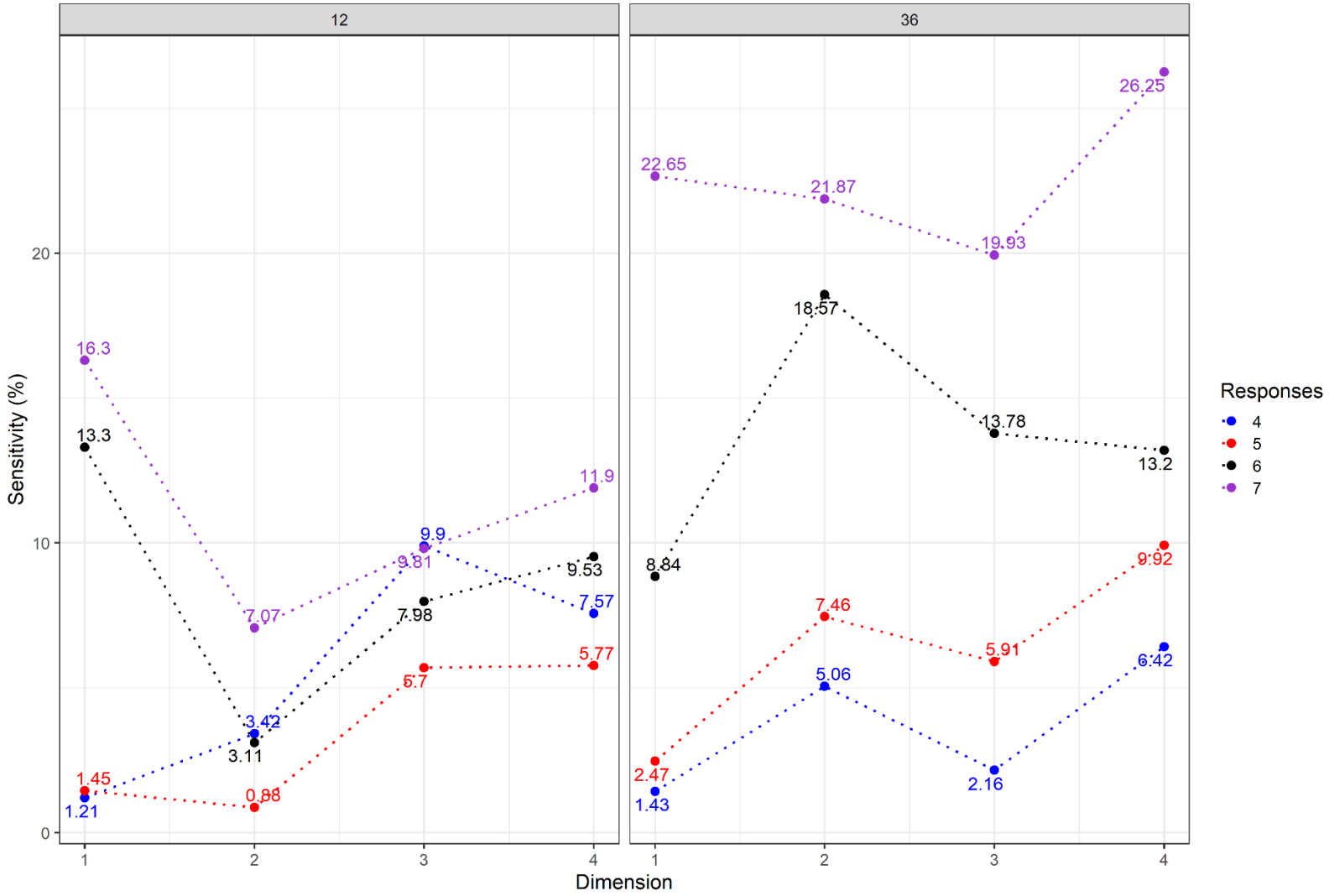


Figure B57

Specificity of H^T_i by Test Length when Applied to Social Desirability Responding

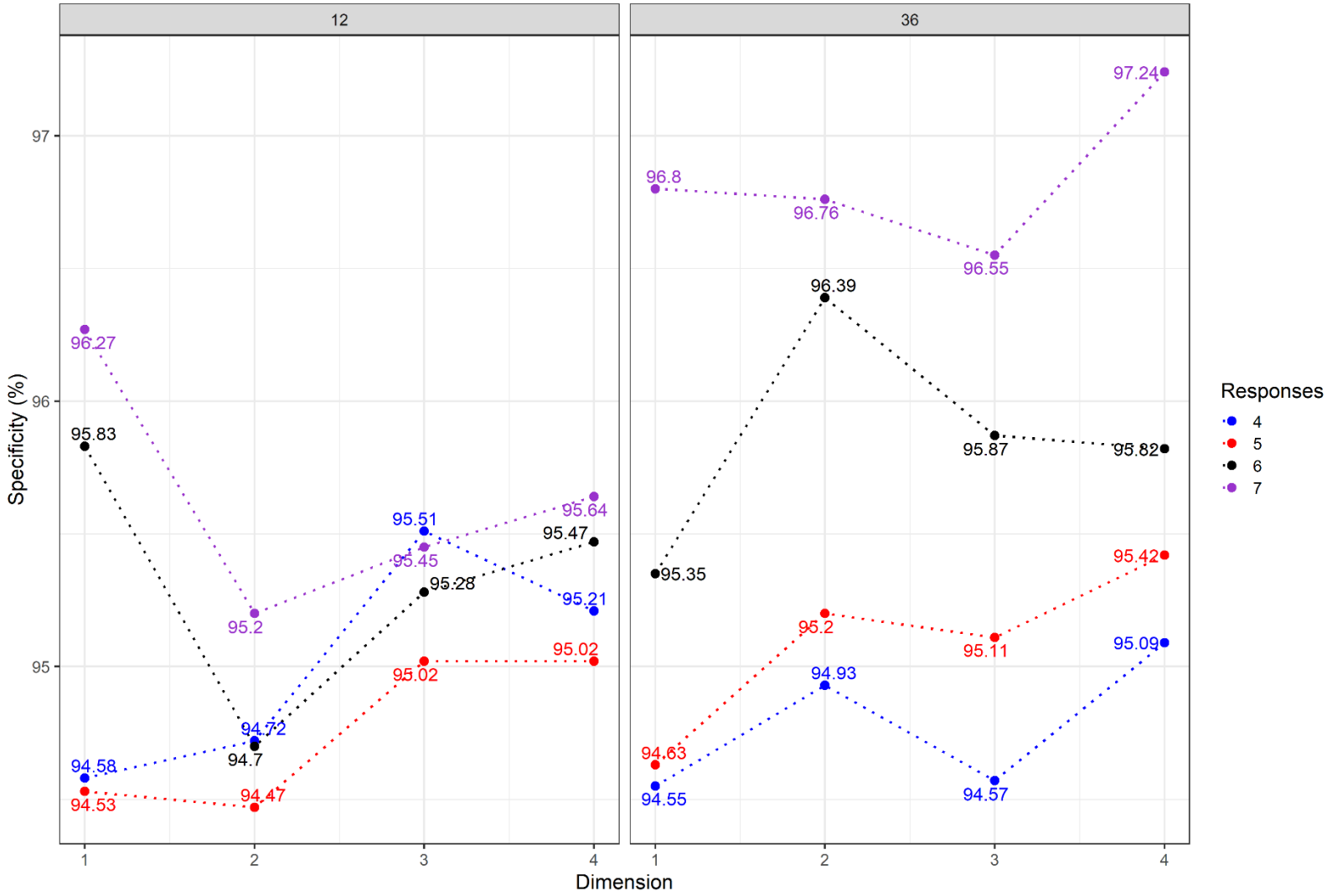


Figure B58

Negative Predictive Value of U3 by Test Length when Applied to Social Desirability Responding

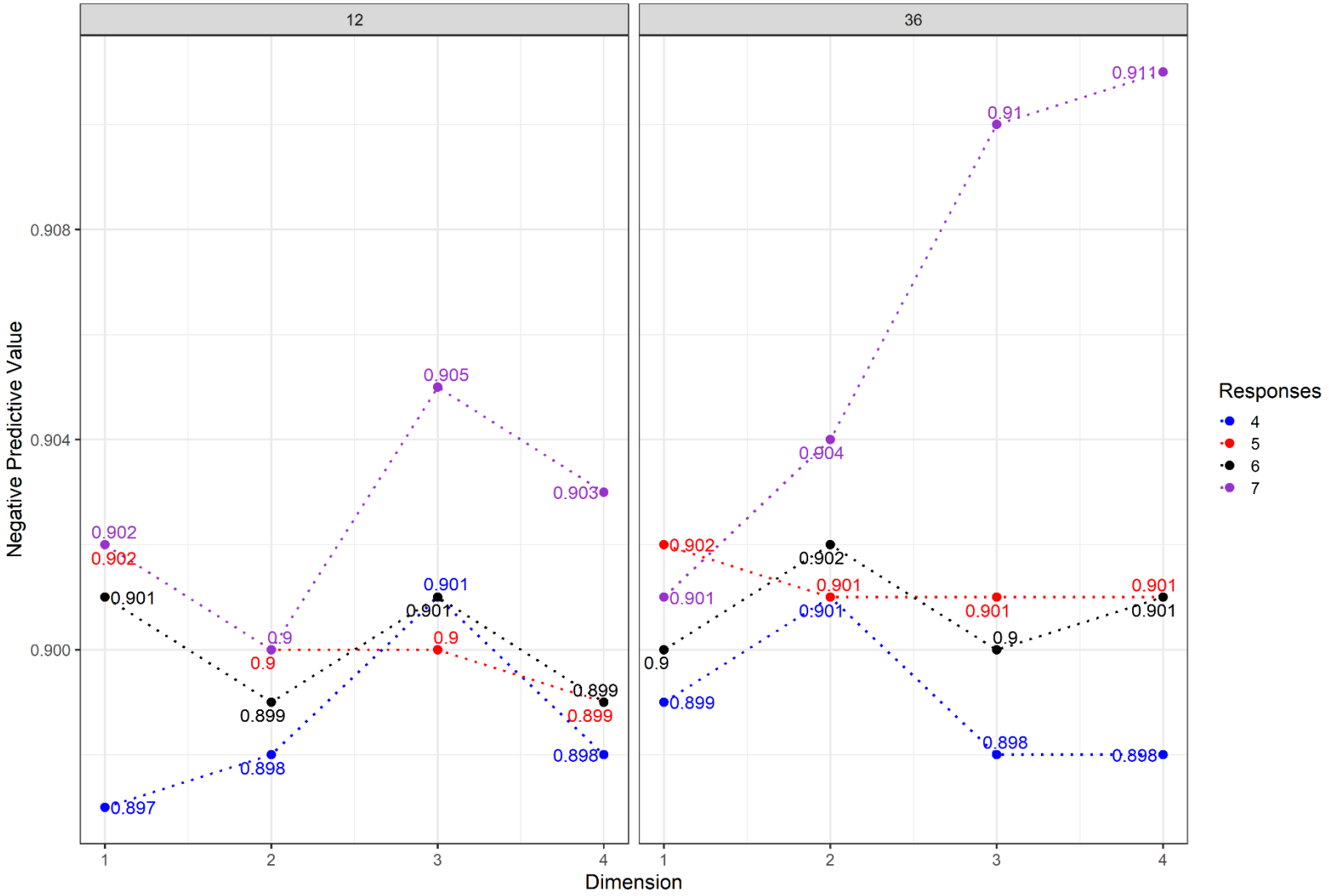


Figure B59

Positive Predictive Value of U3 by Test Length when Applied to Social Desirability Responding

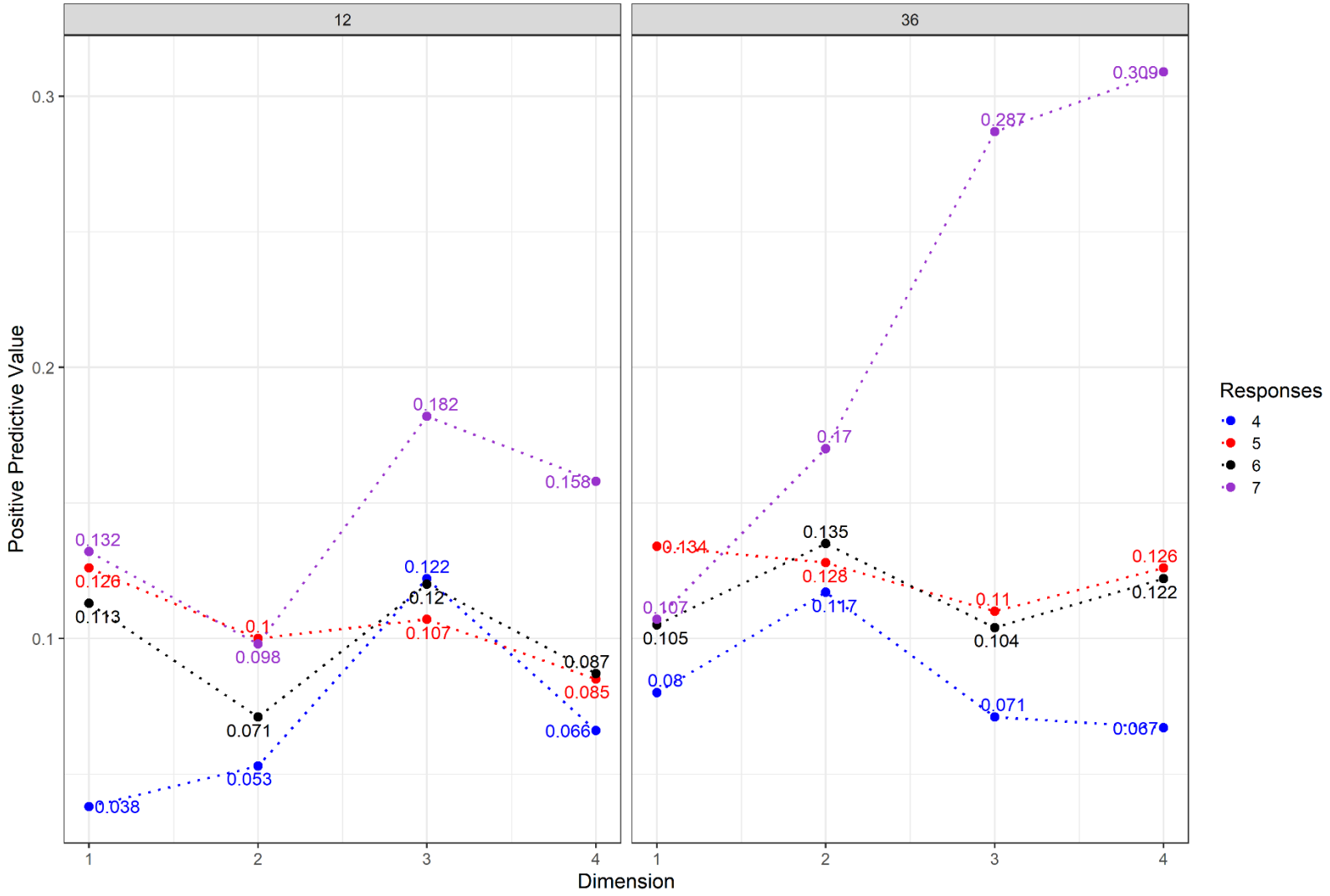


Figure B60

Sensitivity of U3 by Test Length when Applied to Social Desirability Responding

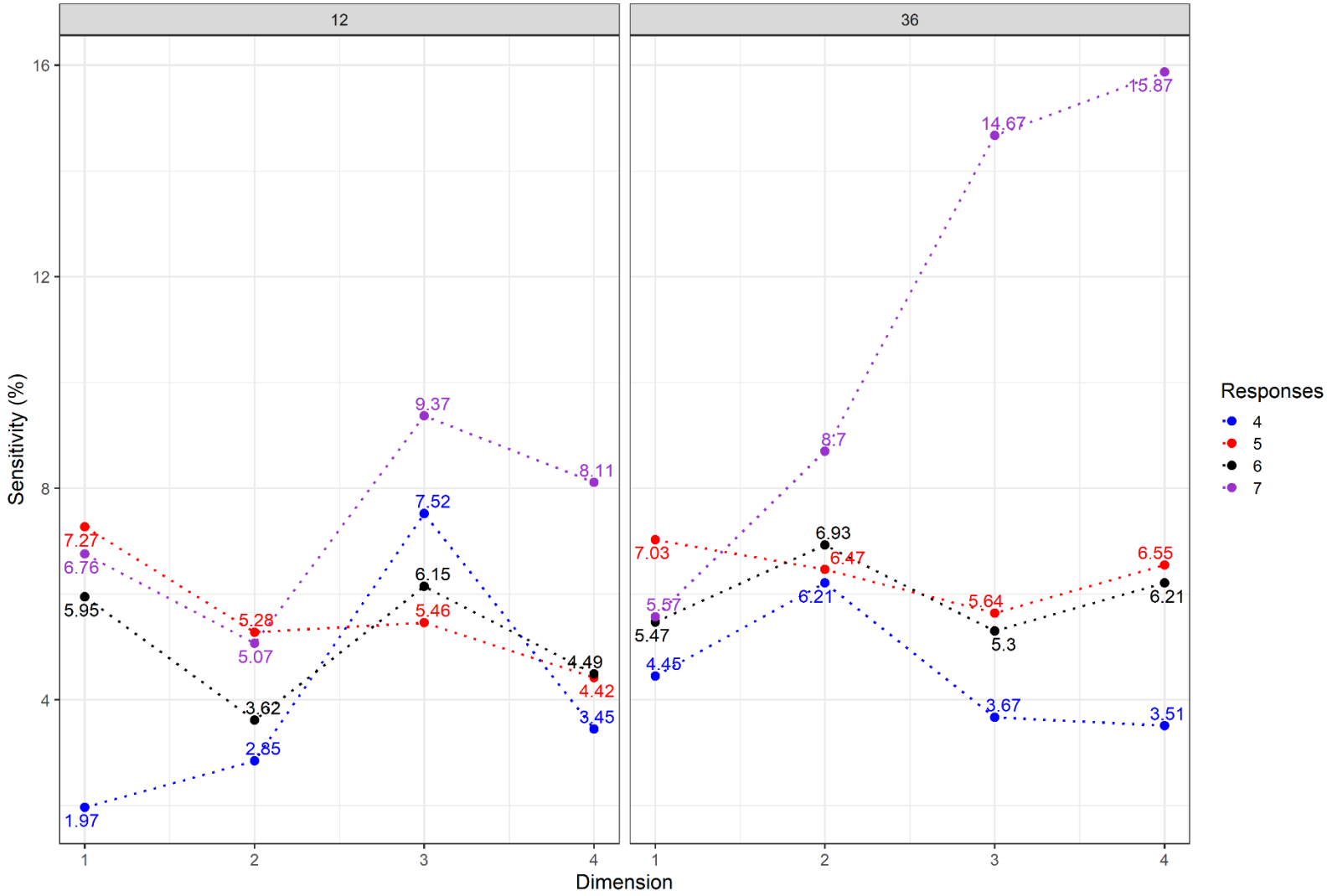


Figure B61

Specificity of H^T_i by Test Length when Applied to Social Desirability Responding

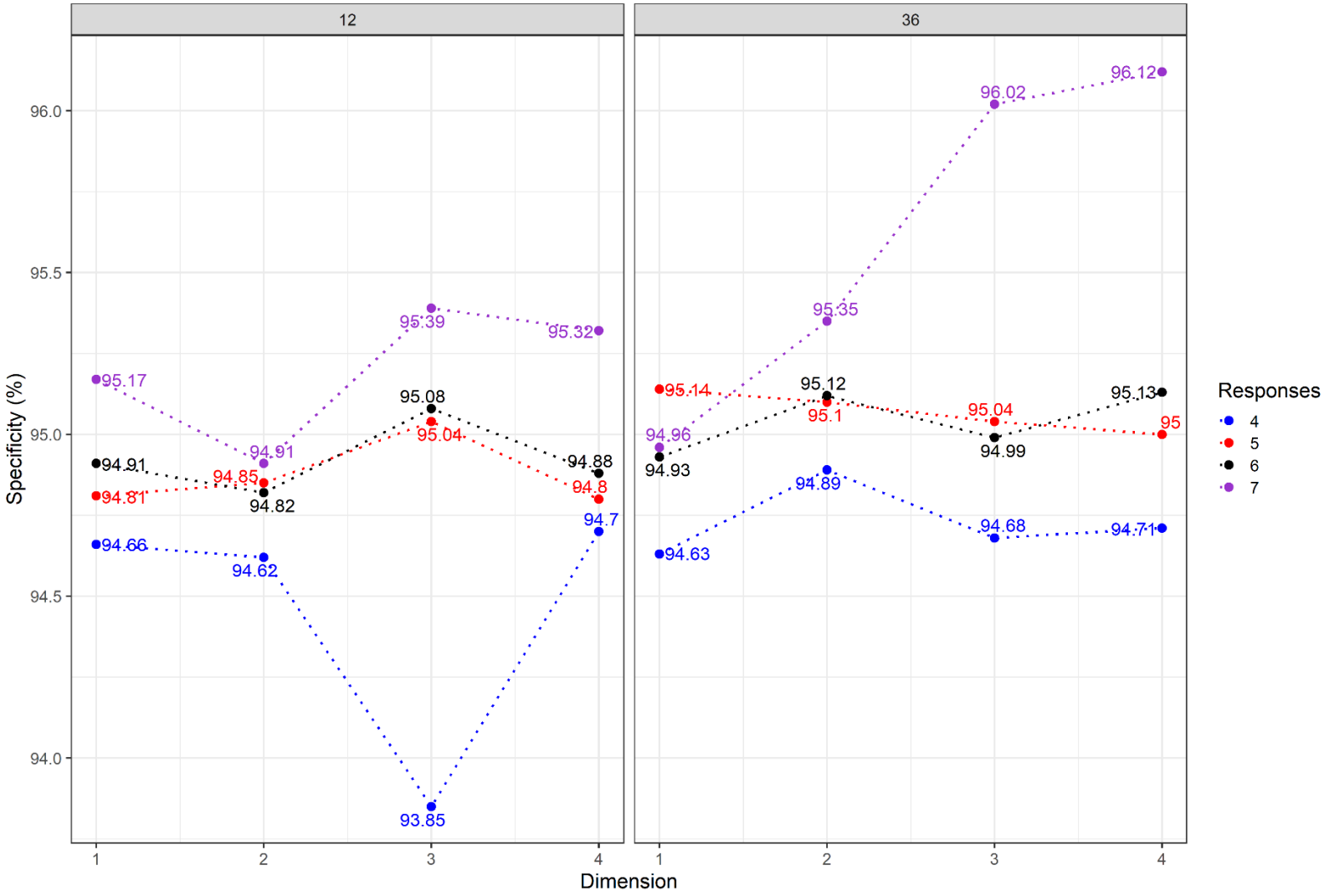


Figure B62

Negative Predictive Value of Guttman Errors by Test Length when Applied to Careless Responding

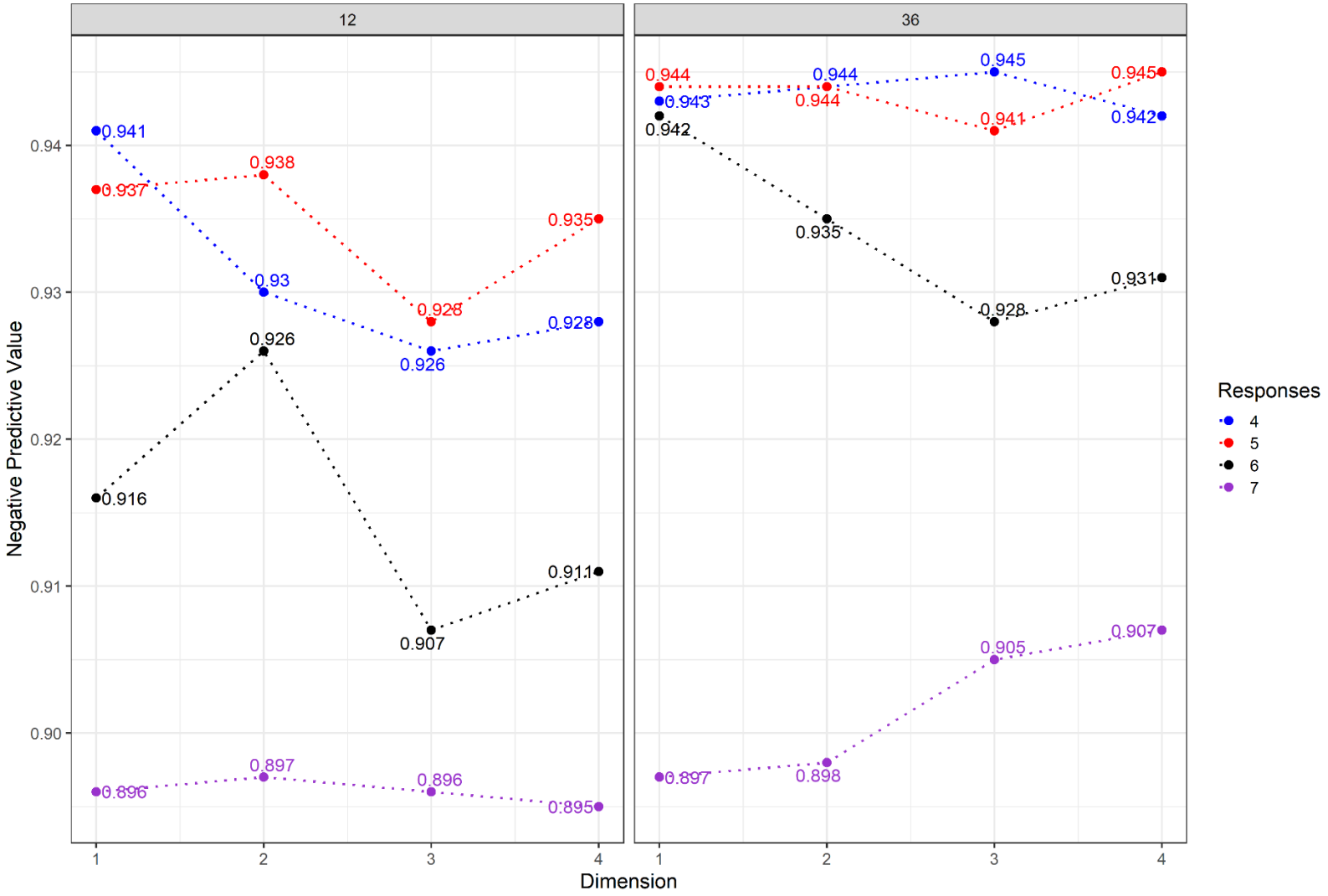


Figure B63

Positive Predictive Value of Guttman Errors by Test Length when Applied to Careless Responding

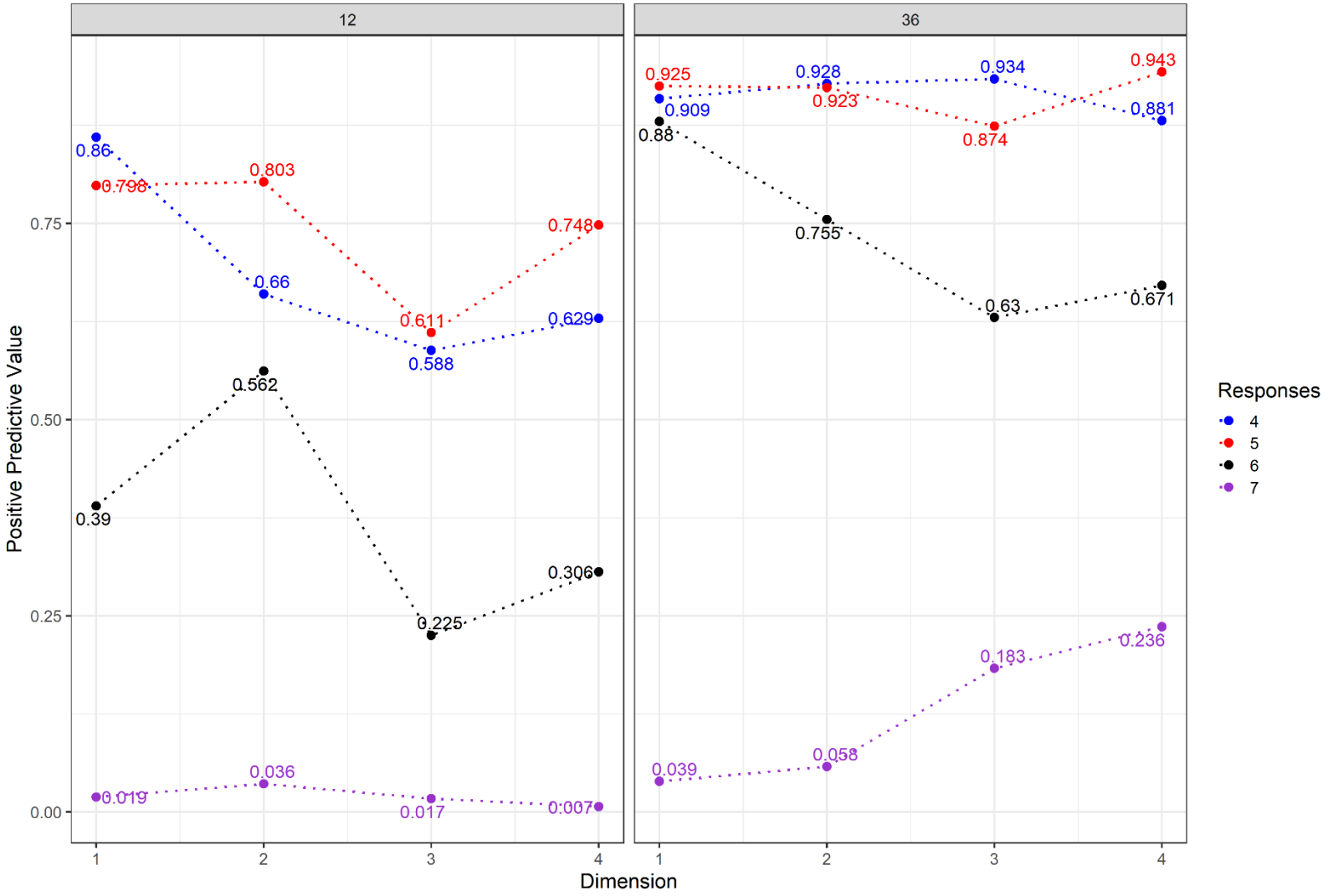


Figure B64

Sensitivity of Guttman Errors by Test Length when Applied to Careless Responding

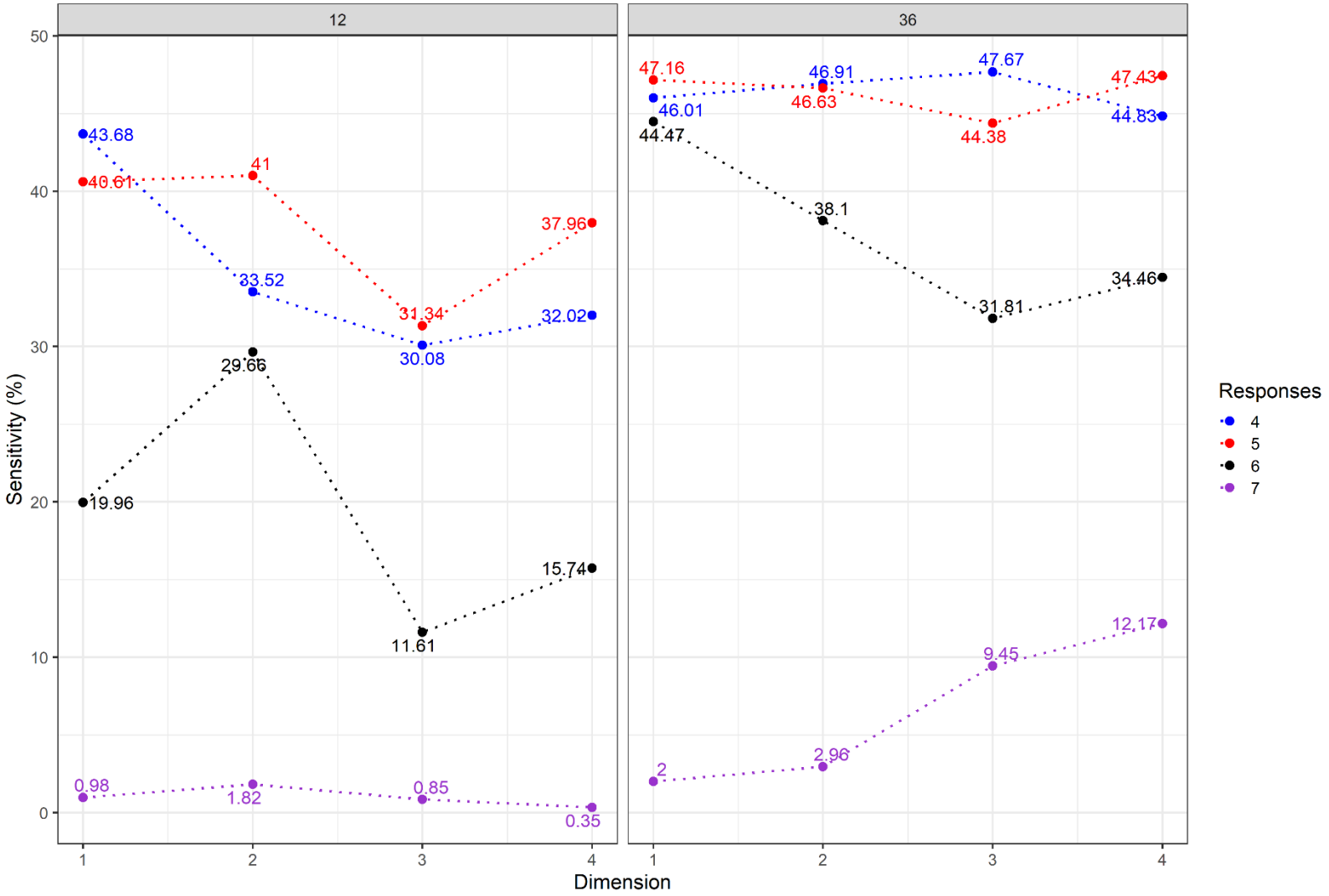


Figure B65

Specificity of Guttman Errors by Test Length when Applied to Careless Responding

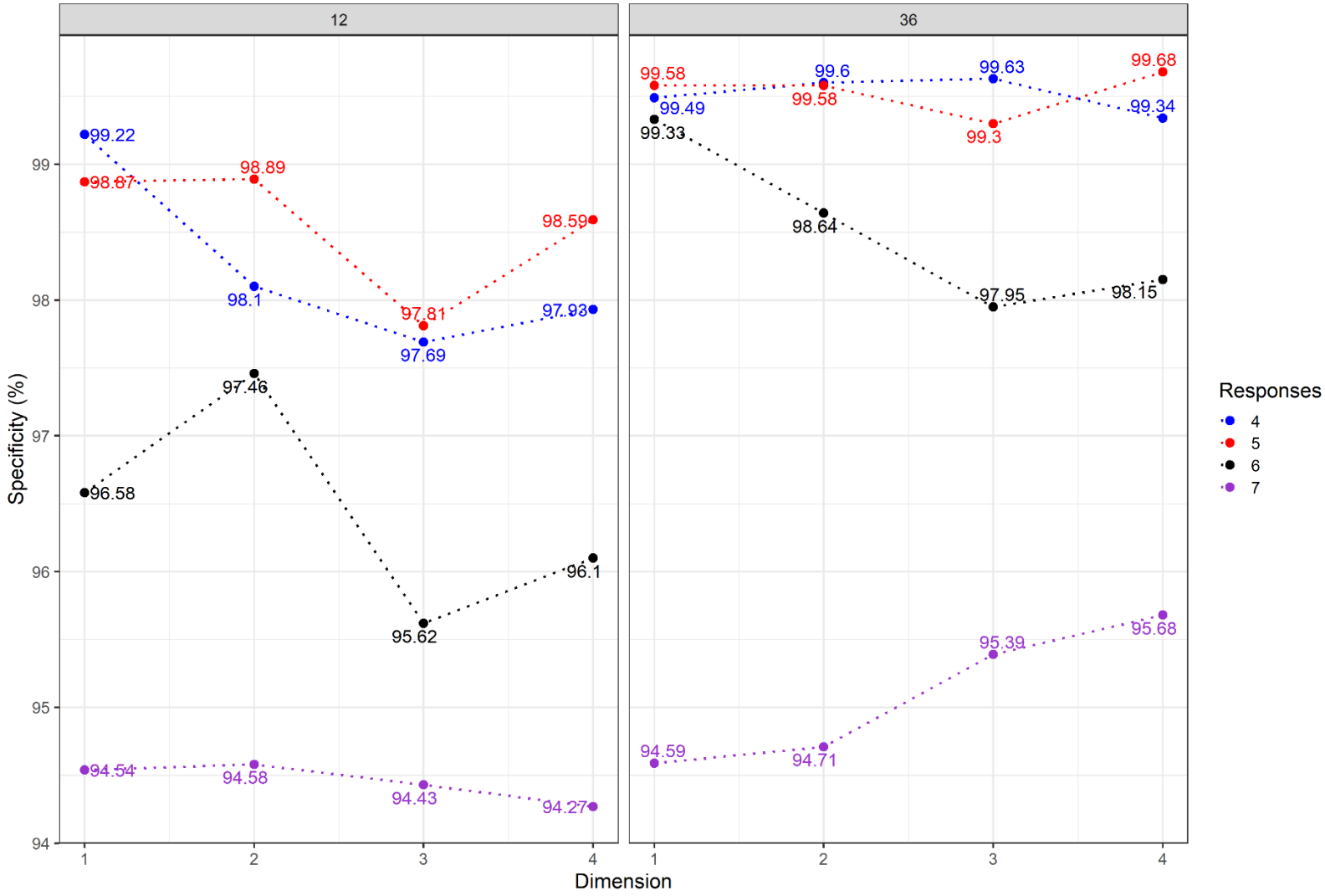


Figure B66

Negative Predictive Value of H^T_i by Test Length when Applied to Careless Responding

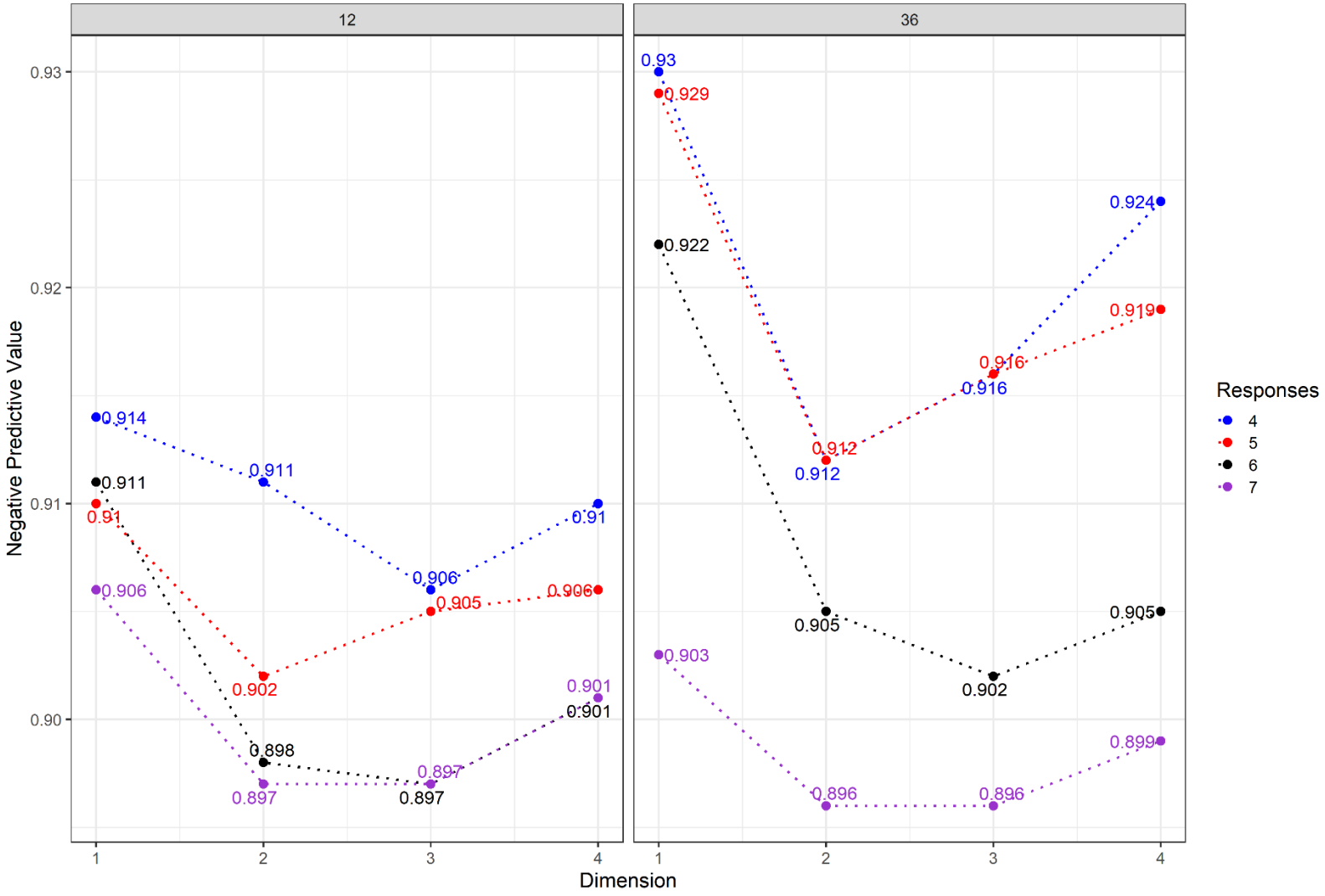


Figure B67

Positive Predictive Value of H^T_i by Test Length when Applied to Careless Responding

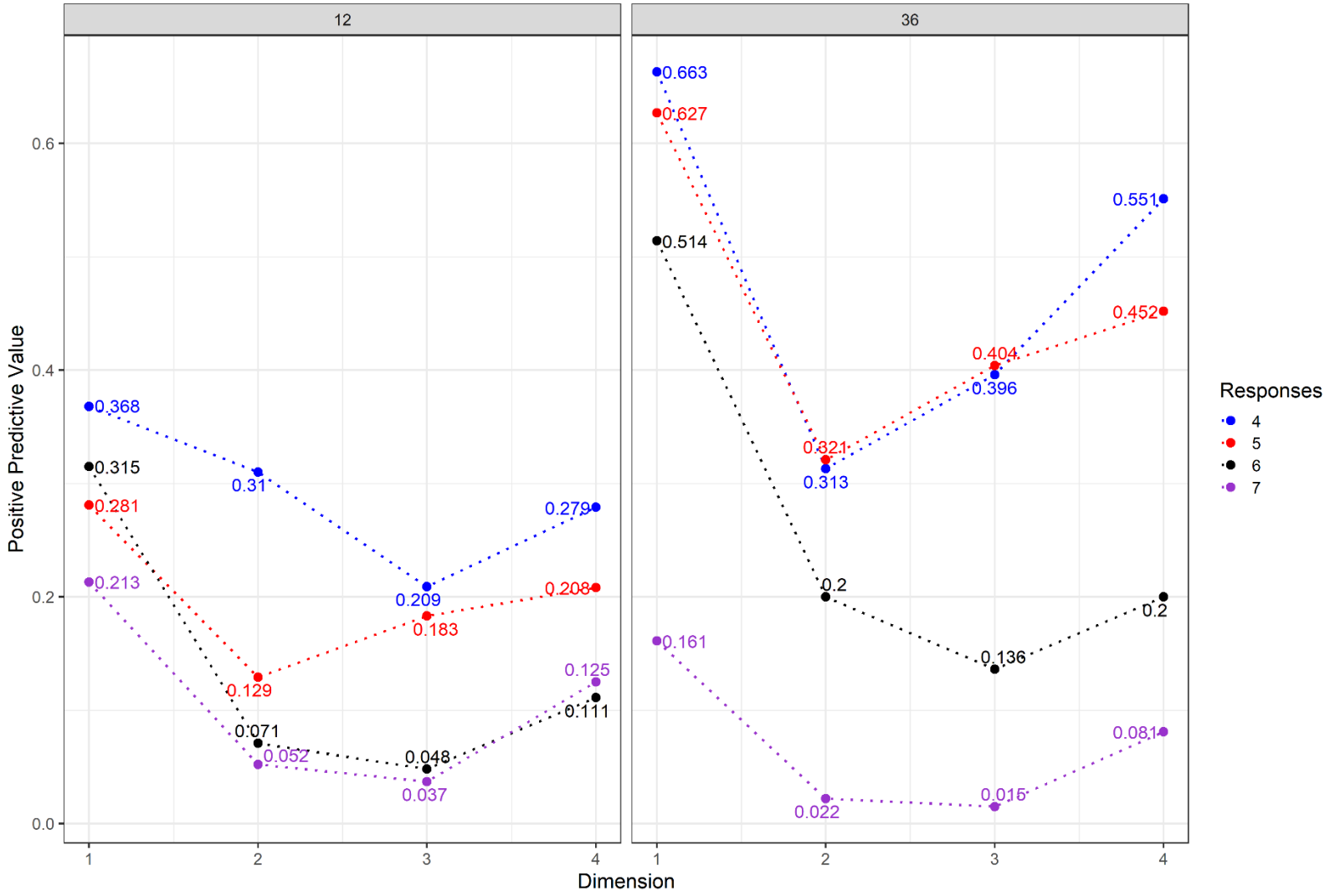


Figure B68

Sensitivity of H^T_i by Test Length when Applied to Careless Responding

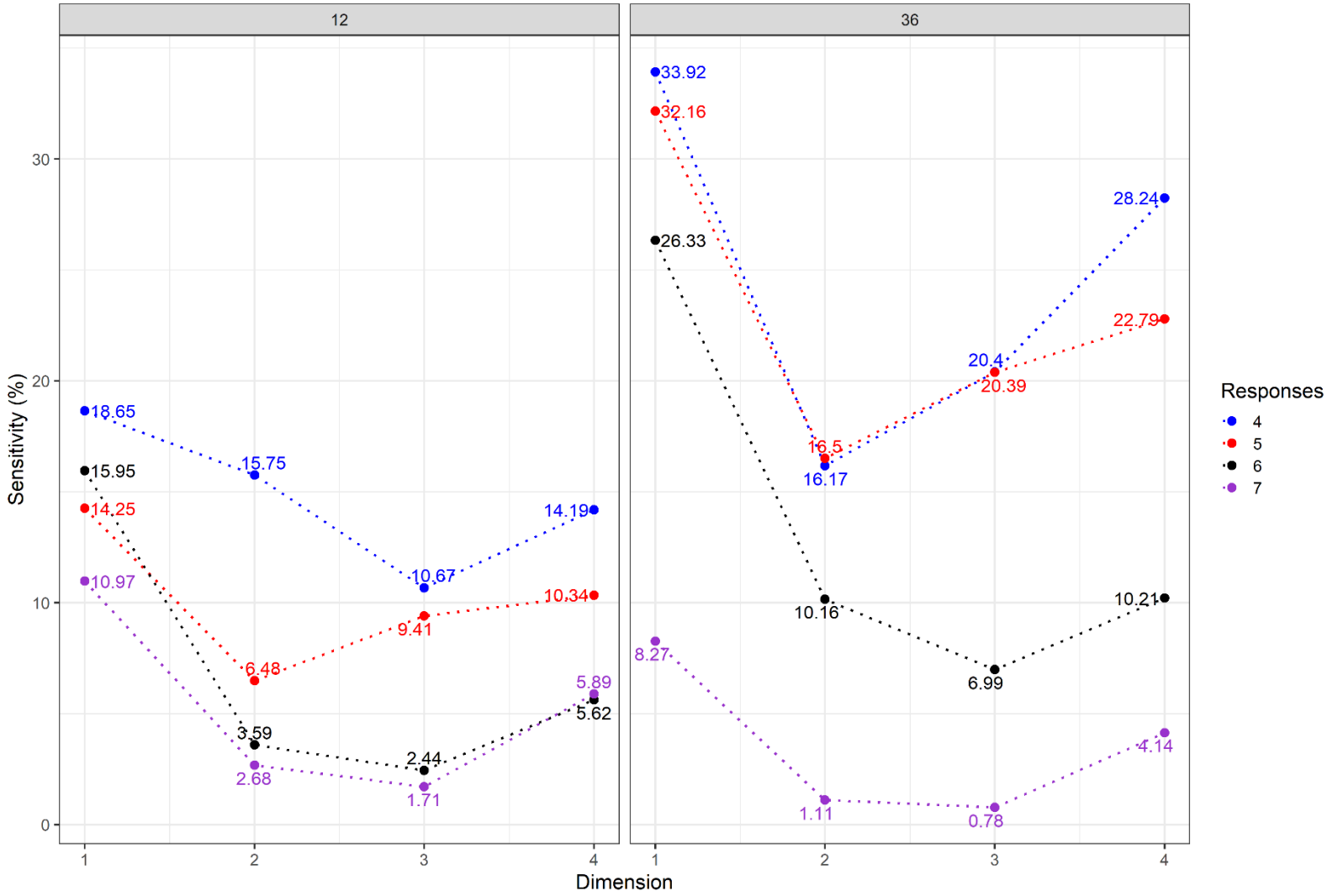


Figure B69

Specificity of H^T_i by Test Length when Applied to Careless Responding

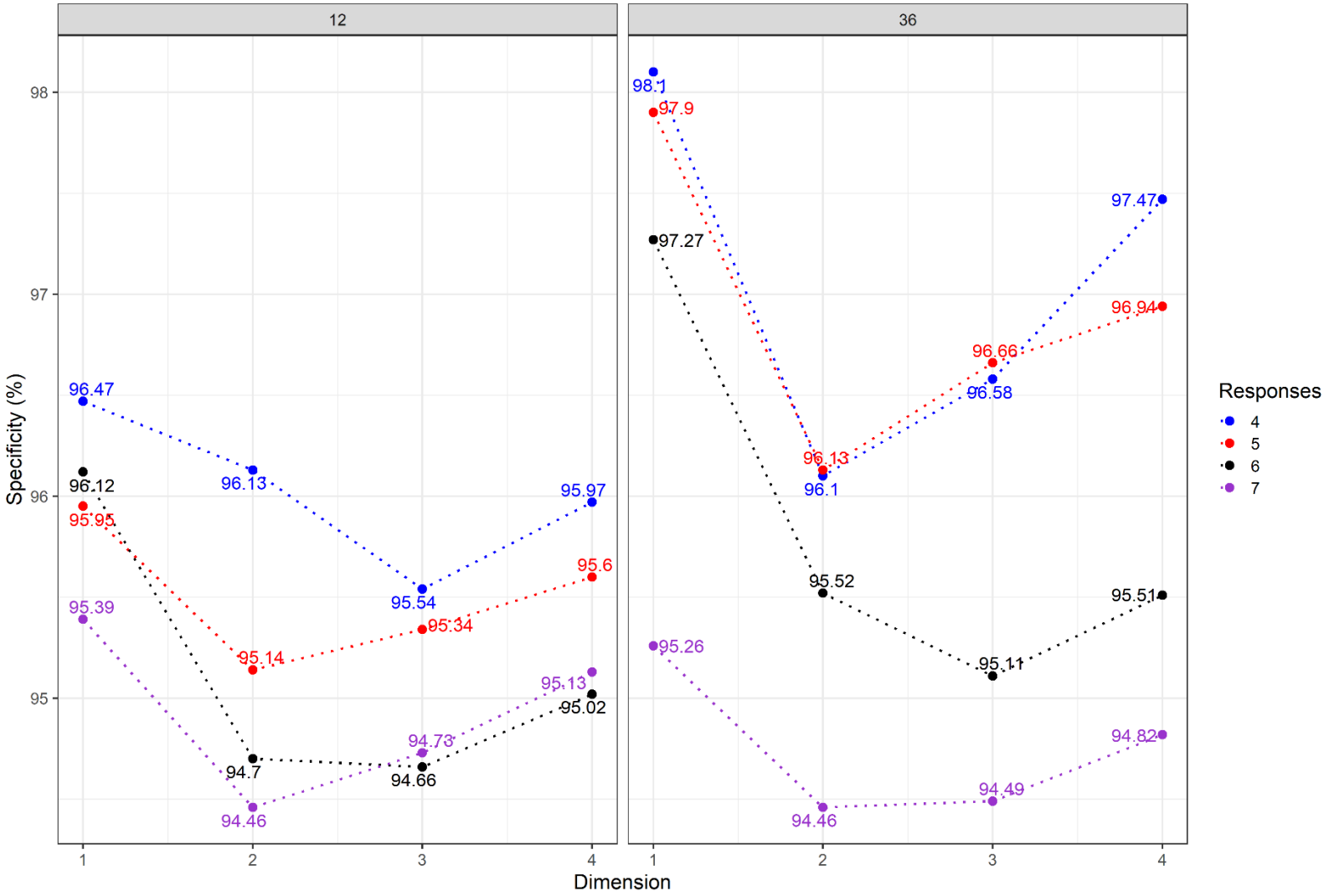


Figure B70

Negative Predictive Value of U3 by Test Length when Applied to Careless Responding

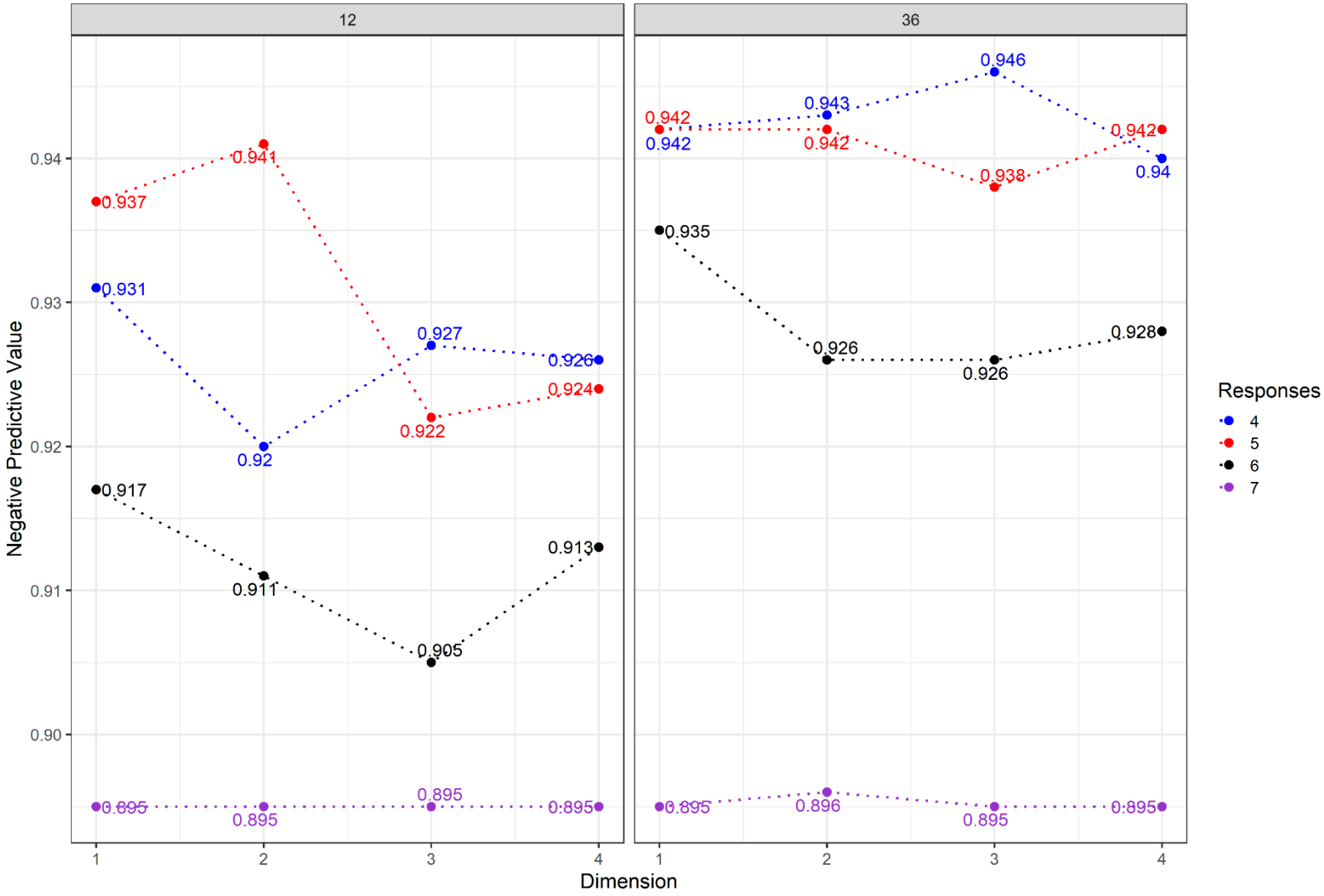


Figure B71

Positive Predictive Value of U3 by Test Length when Applied to Careless Responding

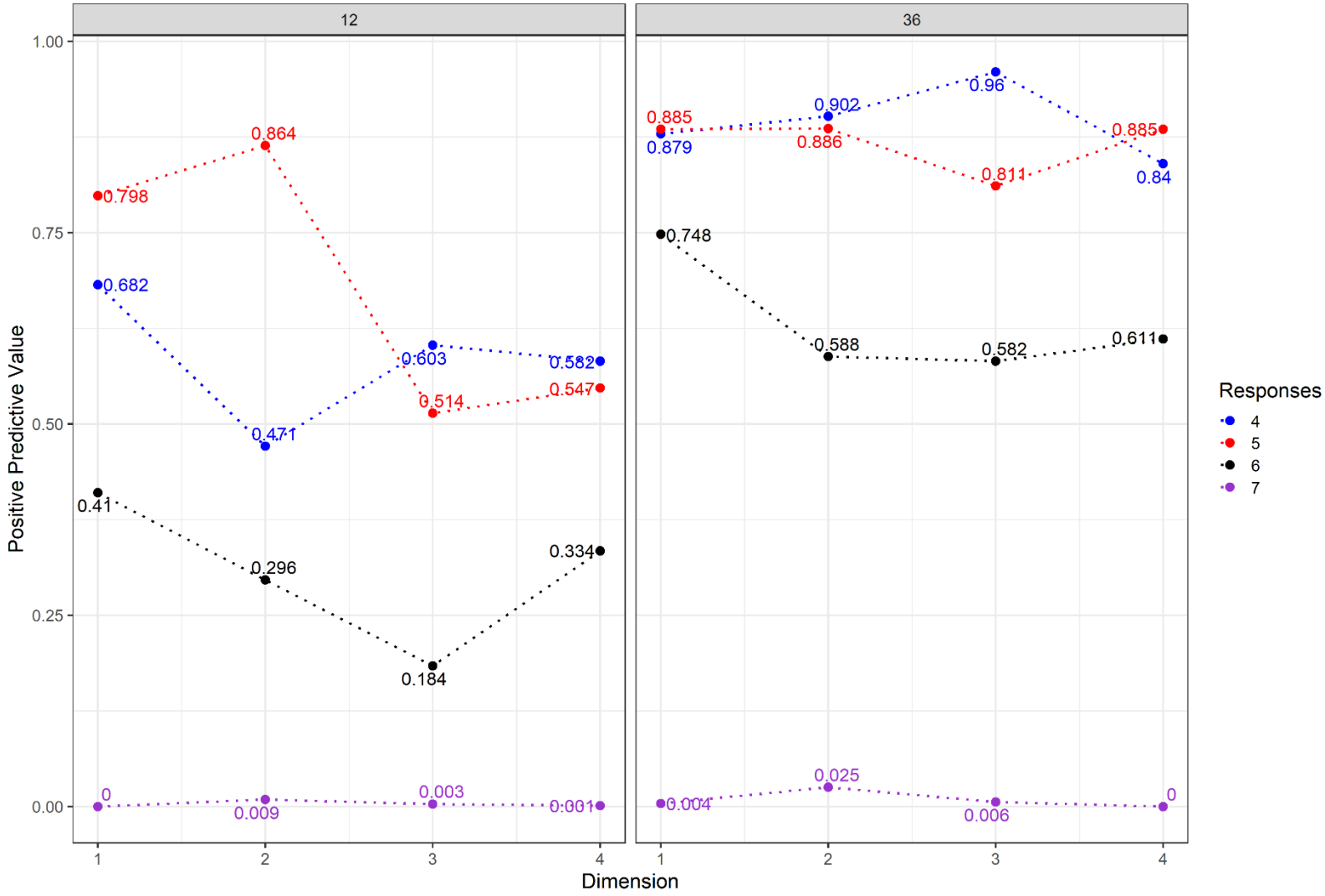


Figure B72

Sensitivity of U3 by Test Length when Applied to Careless Responding

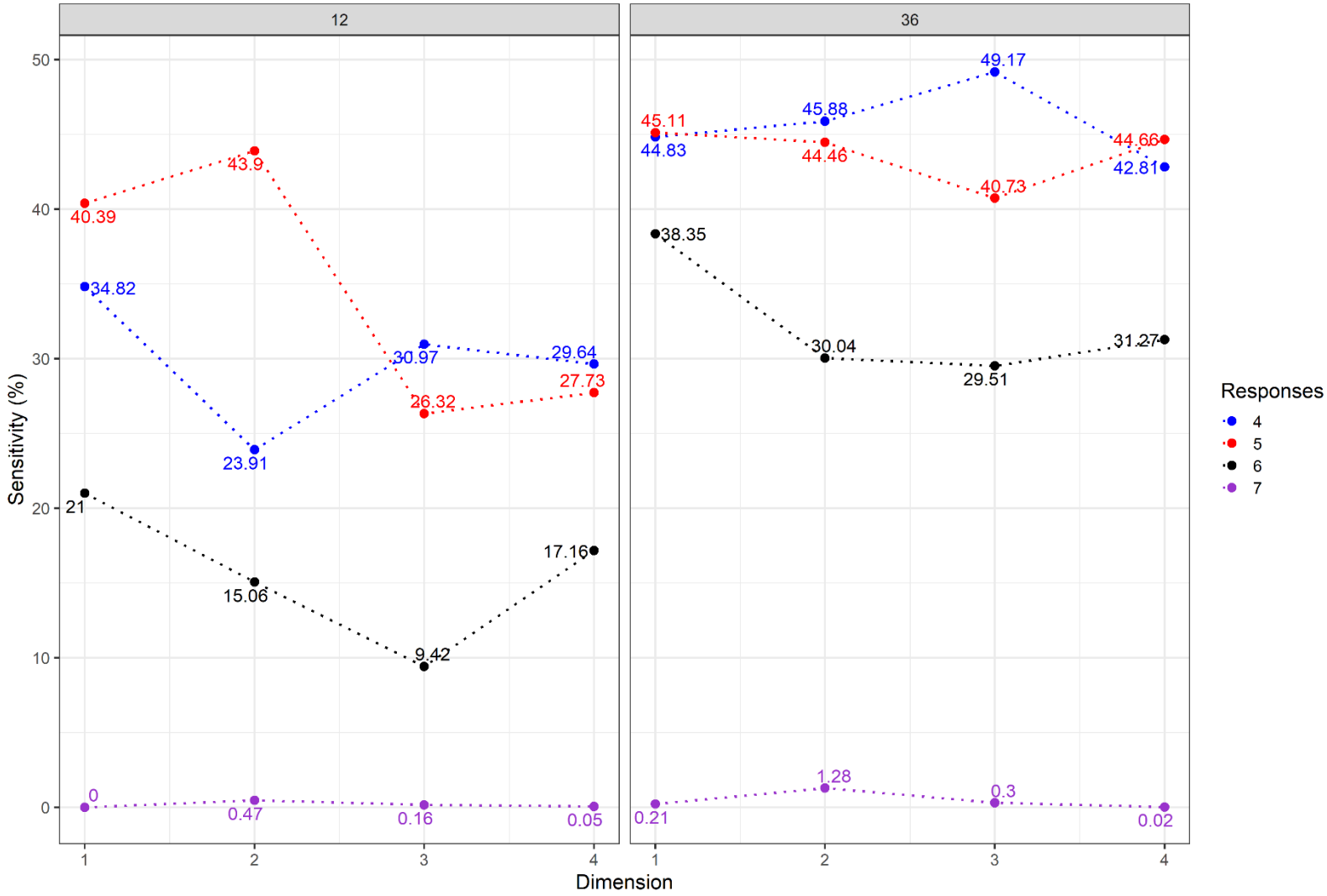


Figure B73

Specificity of U3 by Test Length when Applied to Careless Responding

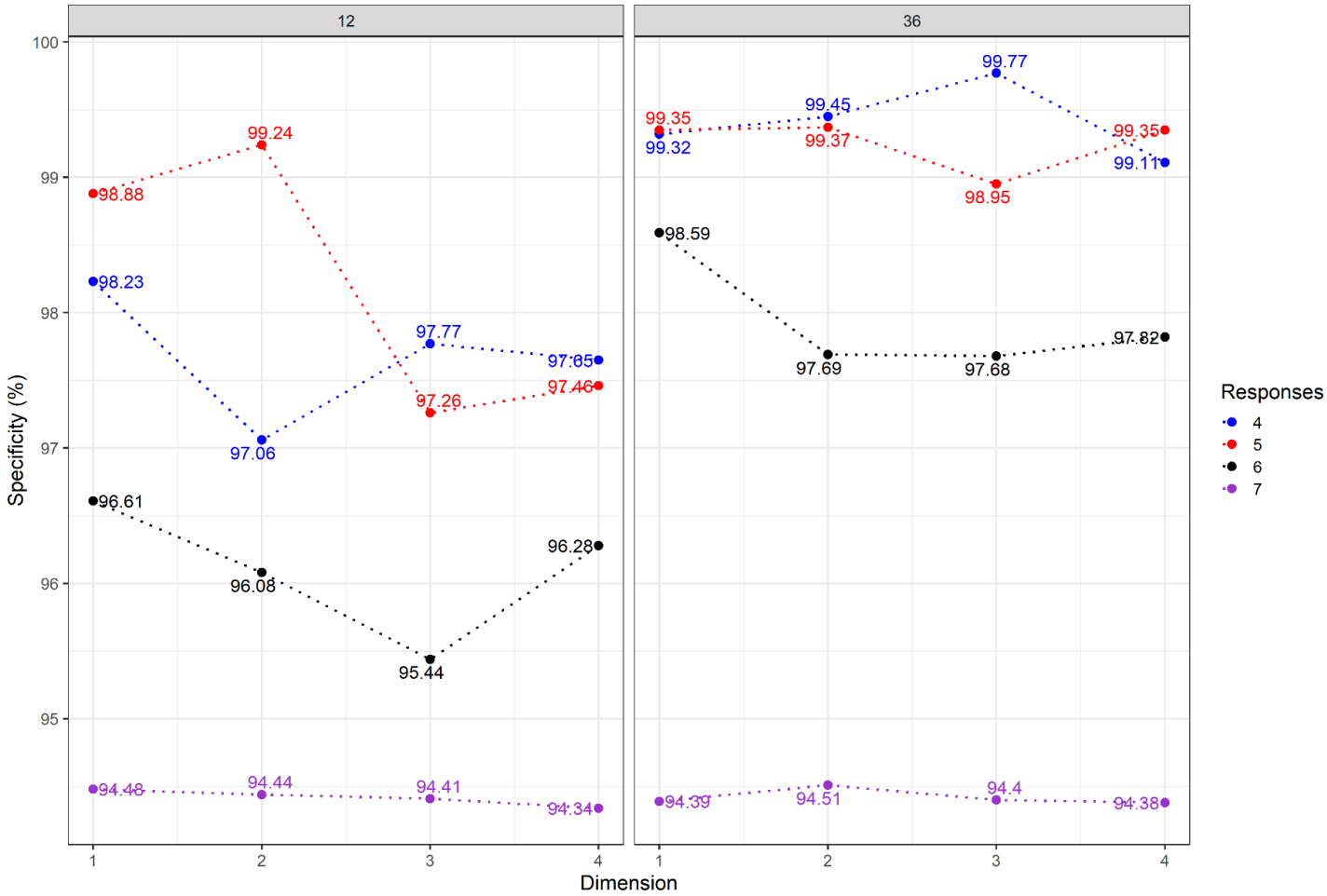


Figure B74

NPV and Specificity of Guttman Errors Aggregated over Aberrant Response Patterns

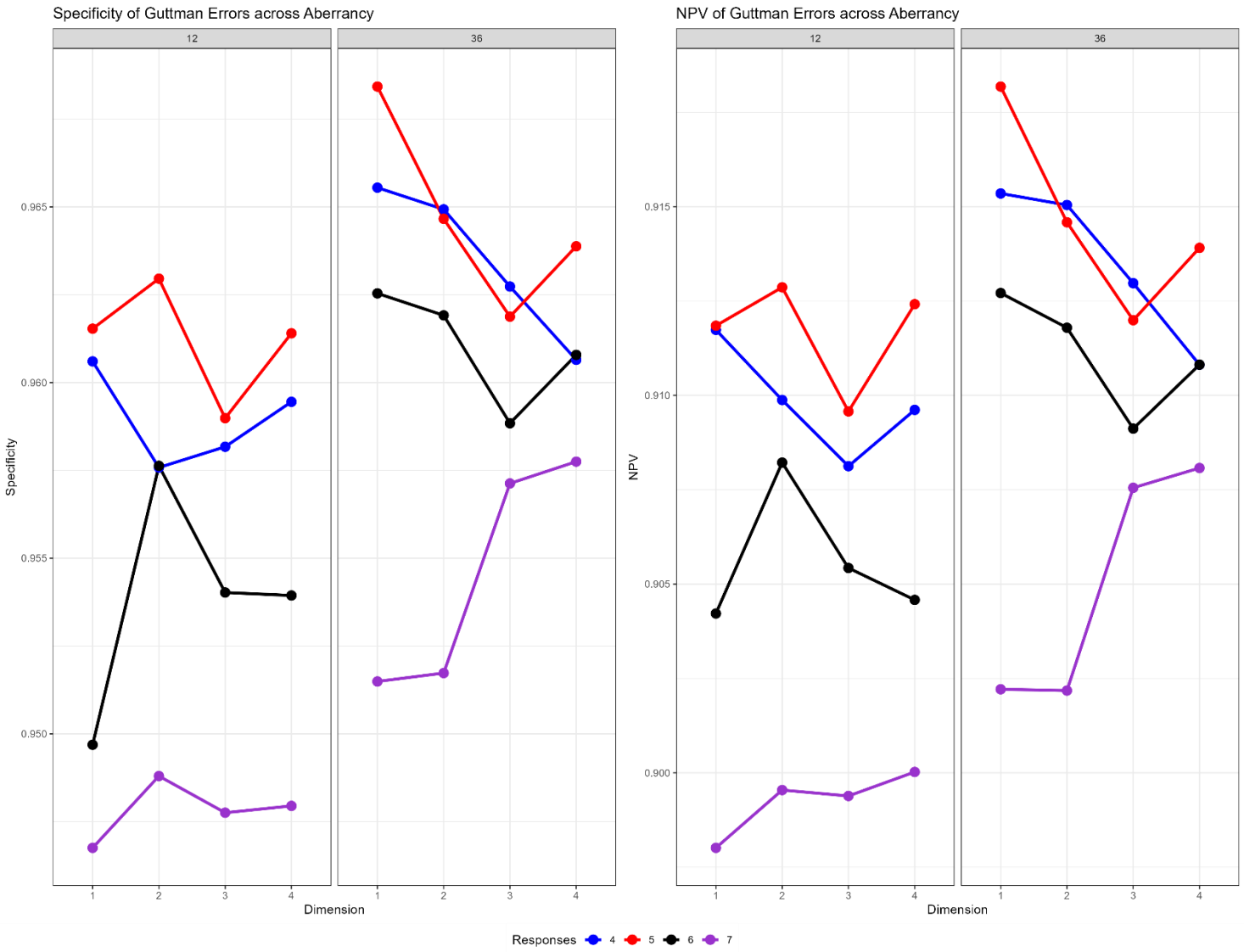


Figure B75

PPV and Sensitivity of Guttman Errors Aggregated over Aberrant Response Patterns

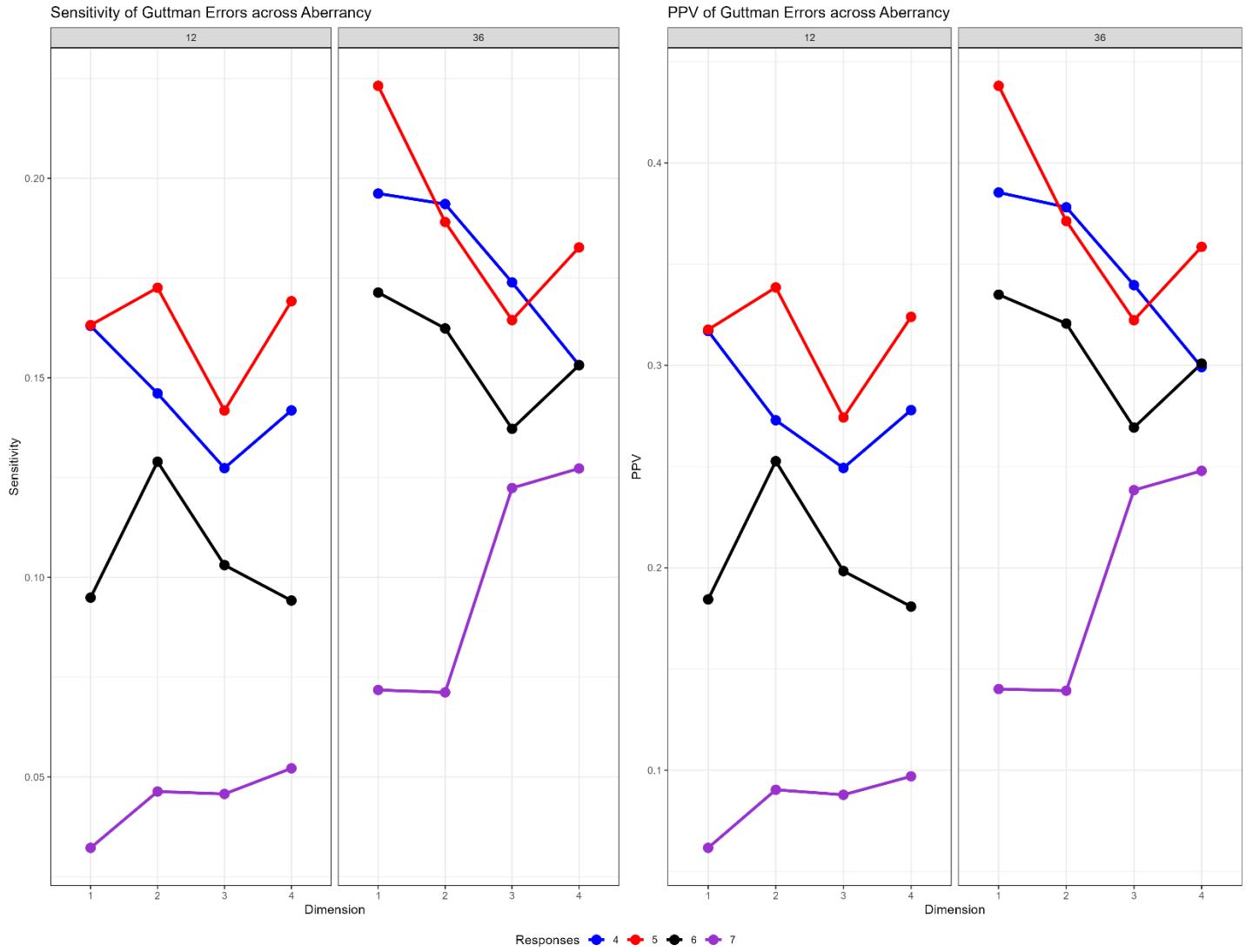


Figure B76

NPV and Specificity of H^T_i Aggregated over Aberrant Response Patterns

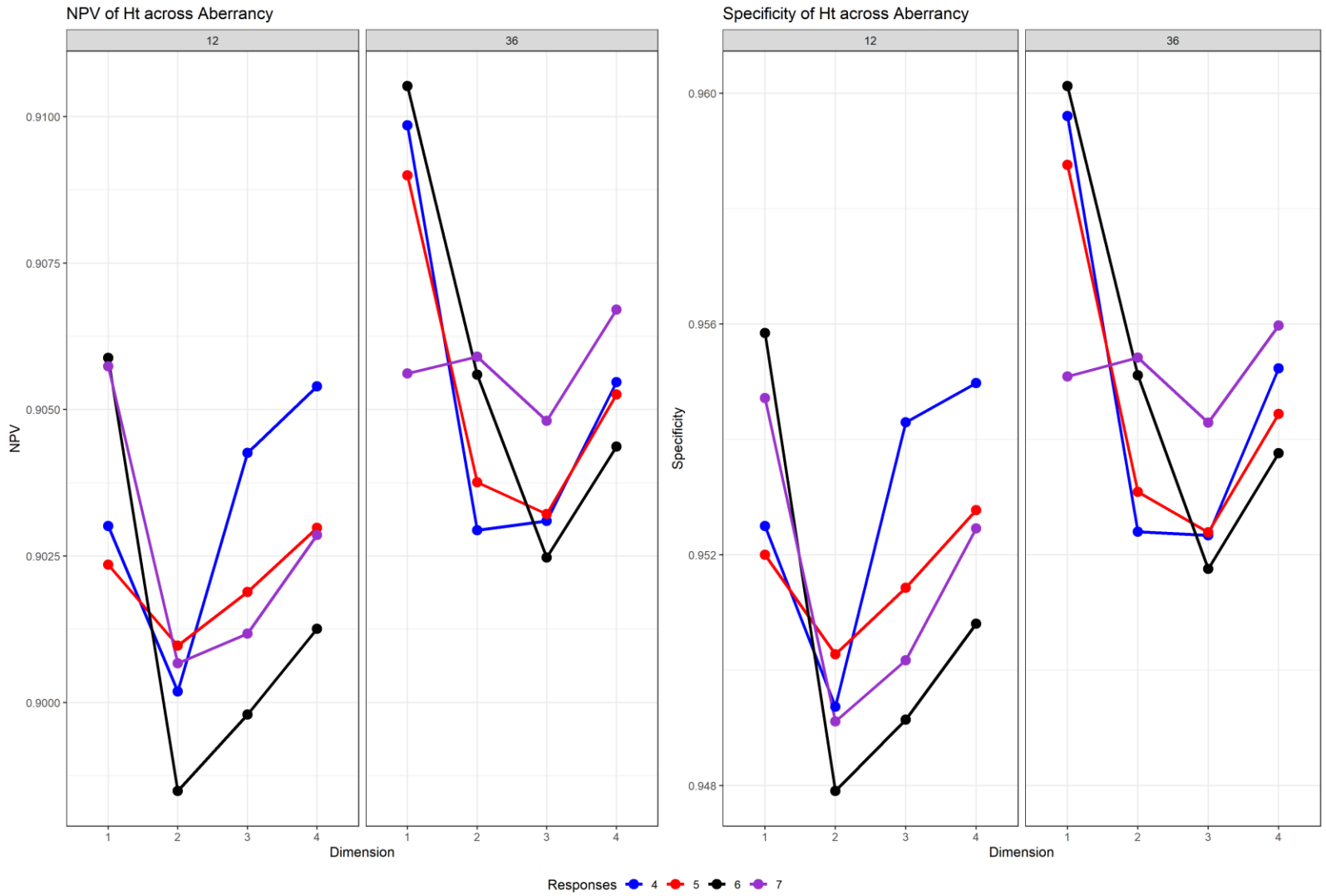


Figure B77

PPV and Sensitivity of H^T_i Aggregated over Aberrant Response Patterns

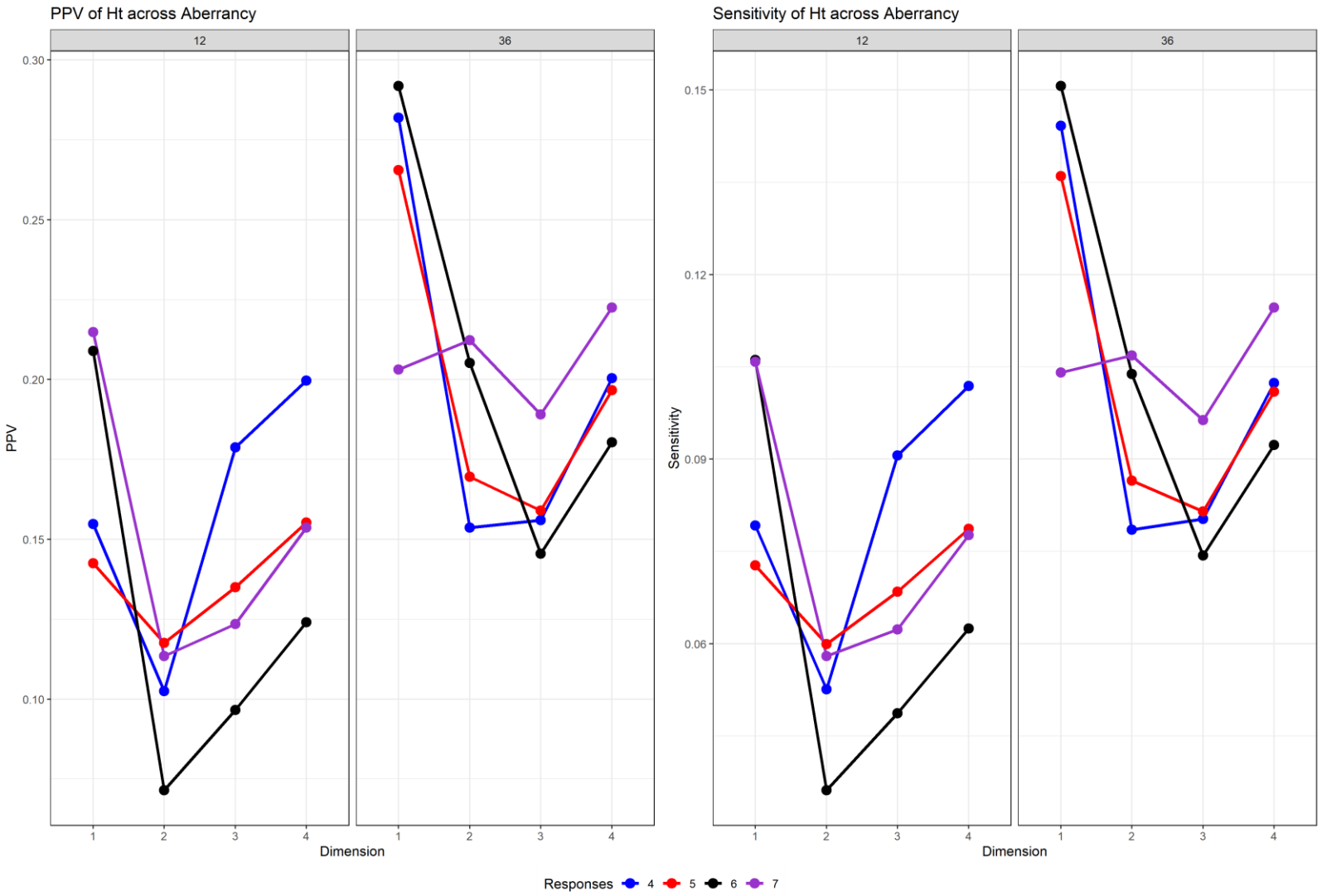


Figure B78

NPV and Specificity of U3 Aggregated over Aberrant Response Patterns

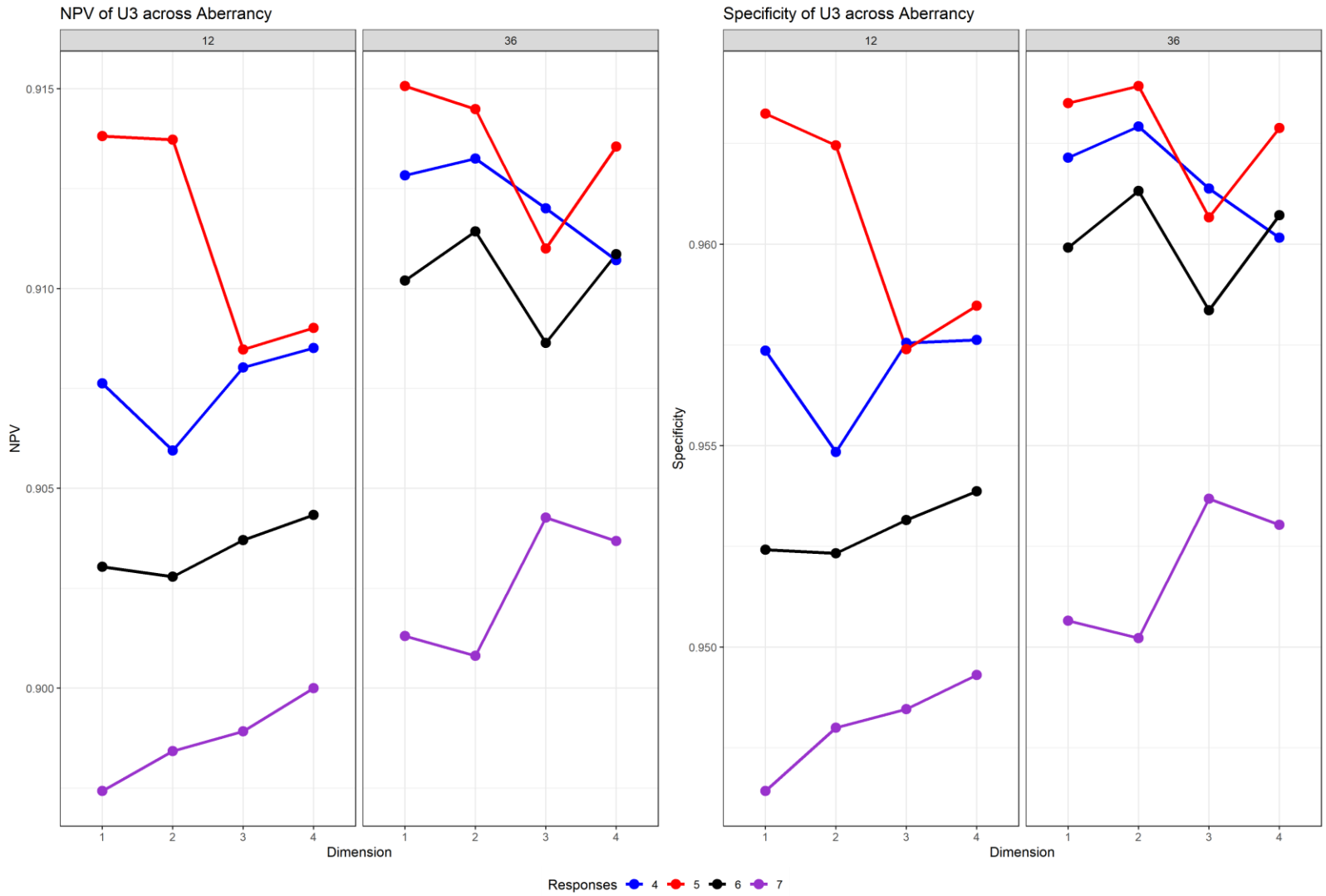
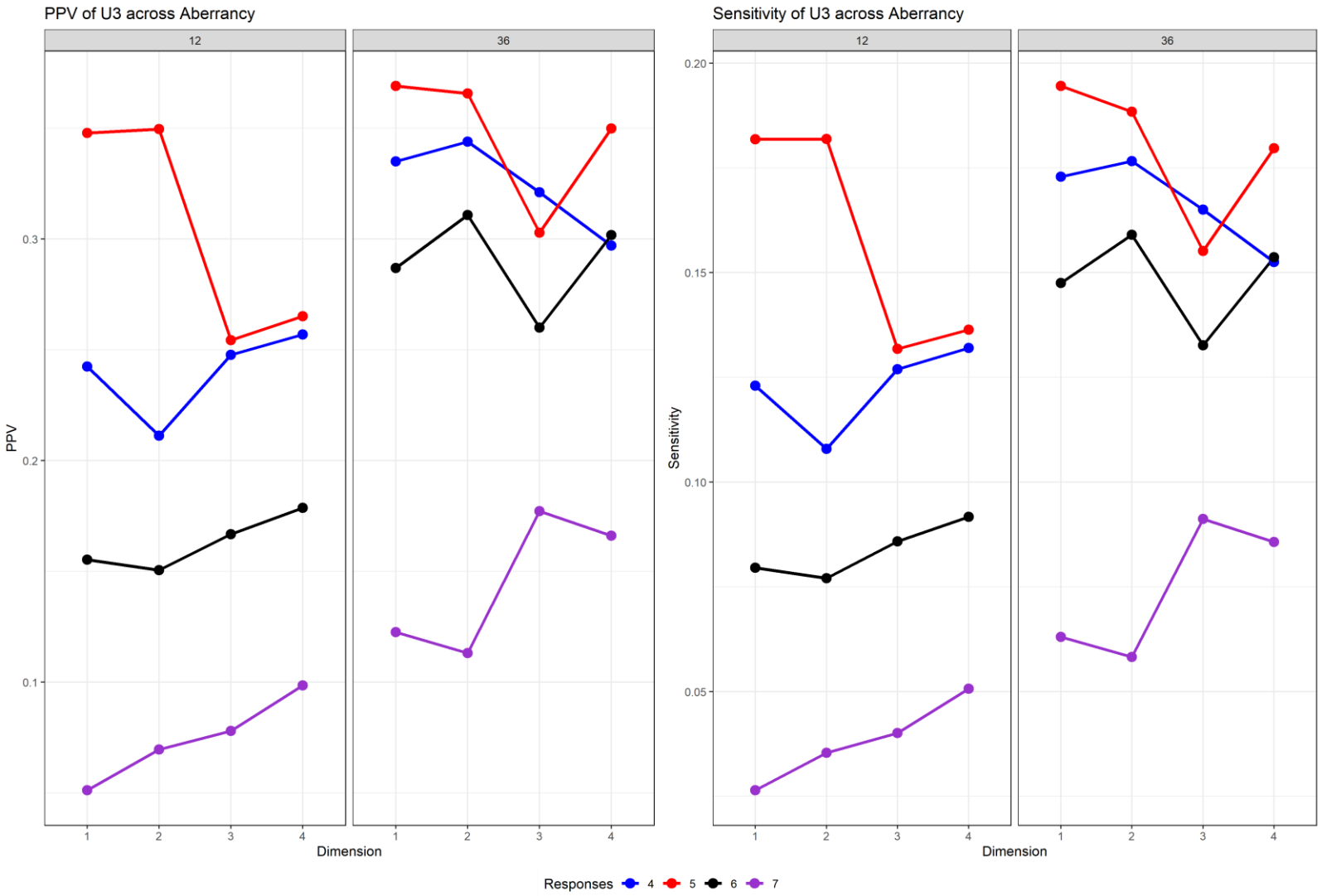


Figure B79

PPV and Sensitivity of U3 Aggregated over Aberrant Response Patterns



APPENDIX C

ANOVA Results for Acquiescence Responding

Table C1

Guttman Errors ANOVA Table for Acquiescence

	Parameter	Sum Sq	df	Mean Sq	F value	Partial Omega Sq
Sensitivity	dim	6.18	3	2.06	4514.87	0.30
	resp	14.35	3	4.78	10490.94	0.50
	leng	5.19	1	5.19	11374.61	0.26
	dim:resp	4.30	9	0.48	1046.67	0.23
	dim:leng	0.92	3	0.31	670.98	0.06
	resp:leng	0.96	3	0.32	700.76	0.06
	dim:resp:leng	4.61	9	0.51	1122.85	0.24
	residual	14.58	31968	0.00		
Specificity	dim	0.13	3	0.04	7919.82	0.43
	resp	0.20	3	0.07	12145.15	0.53
	leng	0.08	1	0.08	14125.70	0.31
	dim:resp	0.07	9	0.01	1319.66	0.27
	dim:leng	0.02	3	0.01	1330.25	0.11
	resp:leng	0.01	3	0.00	840.35	0.07
	dim:resp:leng	0.07	9	0.01	1436.71	0.29
	residual	0.18	31968	0.00		
NPV	dim	0.07	3	0.02	5898.14	0.36
	resp	0.16	3	0.05	13005.26	0.55
	leng	0.06	1	0.06	14269.94	0.31
	dim:resp	0.05	9	0.01	1277.13	0.26
	dim:leng	0.01	3	0.00	836.01	0.07
	resp:leng	0.01	3	0.00	836.83	0.07
	dim:resp:leng	0.05	9	0.01	1369.83	0.28
	residual	0.13	31968	0.00		
PPV	dim	23.31	3	7.77	5996.64	0.36
	resp	54.57	3	18.19	14039.15	0.57
	leng	19.30	1	19.30	14899.40	0.32
	dim:resp	15.81	9	1.76	1356.01	0.28
	dim:leng	3.44	3	1.15	885.82	0.08
	resp:leng	3.68	3	1.23	946.52	0.08

	dim:resp:leng	17.61	9	1.96	1510.45	0.30
	residual	41.42	31968	0.00		

Table C2

H^T_i ANOVA Tables for Acquiescence

	Parameter	Sum Sq	df	Mean Sq	F value	Partial Omega Sq
Sensitivity	dim	1.92	3	0.64	1078.32	0.09
	resp	57.82	3	19.27	32439.50	0.75
	leng	1.03	1	1.03	1732.85	0.05
	dim:resp	2.92	9	0.32	545.84	0.13
	dim:leng	7.37	3	2.46	4133.80	0.28
	resp:leng	3.43	3	1.14	1924.43	0.15
	dim:resp:leng	4.06	9	0.45	758.33	0.18
	residual	18.99	31968	0.00		
Specificity	dim	0.02	3	0.01	238.45	0.02
	resp	0.63	3	0.21	6230.53	0.37
	leng	0.01	1	0.01	315.64	0.01
	dim:resp	0.03	9	0.00	100.92	0.03
	dim:leng	0.09	3	0.03	910.05	0.08
	resp:leng	0.05	3	0.02	512.65	0.05
	dim:resp:leng	0.04	9	0.00	139.74	0.04
	residual	1.09	31968	0.00		
NPV	dim	0.02	3	0.01	1196.40	0.10
	resp	0.63	3	0.21	35388.84	0.77
	leng	0.01	1	0.01	1881.34	0.06
	dim:resp	0.03	9	0.00	592.39	0.14
	dim:leng	0.08	3	0.03	4561.87	0.30
	resp:leng	0.04	3	0.01	2165.90	0.17
	dim:resp:leng	0.04	9	0.00	823.53	0.19
	residual	0.19	31968	0.00		
PPV	dim	6.85	3	2.28	1007.67	0.09
	resp	227.37	3	75.79	33441.25	0.76
	leng	4.09	1	4.09	1802.87	0.05
	dim:resp	11.16	9	1.24	547.24	0.13
	dim:leng	30.84	3	10.28	4535.92	0.30
	resp:leng	13.82	3	4.61	2032.97	0.16

	dim:resp:leng	16.00	9	1.78	784.44	0.18
	residual	72.45	31968	0.00		

Table C3

U3 ANOVA Tables for Acquiescence

	Parameter	Sum Sq	df	Mean Sq	F value	Partial Omega Sq
Sensitivity	dim	1.76	3	0.59	382.47	0.03
	resp	16.83	3	5.61	3658.98	0.26
	leng	5.97	1	5.97	3894.01	0.11
	dim:resp	4.79	9	0.53	347.22	0.09
	dim:leng	2.21	3	0.74	480.23	0.04
	resp:leng	1.21	3	0.40	262.42	0.02
	dim:resp:leng	1.70	9	0.19	123.01	0.03
	residual	49.02	31968	0.00		
Specificity	dim	0.01	3	0.00	69.35	0.01
	resp	0.04	3	0.01	198.07	0.02
	leng	0.06	1	0.06	922.05	0.03
	dim:resp	0.01	9	0.00	12.92	0.00
	dim:leng	0.02	3	0.01	105.30	0.01
	resp:leng	0.02	3	0.01	101.11	0.01
	dim:resp:leng	0.01	9	0.00	16.51	0.00
	residual	2.23	31968	0.00		
NPV	dim	0.02	3	0.01	561.93	0.05
	resp	0.17	3	0.06	5034.65	0.32
	leng	0.07	1	0.07	5917.02	0.16
	dim:resp	0.05	9	0.01	471.81	0.12
	dim:leng	0.02	3	0.01	708.51	0.06
	resp:leng	0.01	3	0.00	422.52	0.04
	dim:resp:leng	0.02	9	0.00	159.60	0.04
	residual	0.36	31968	0.00		
PPV	dim	5.82	3	1.94	1014.48	0.09
	resp	41.73	3	13.91	7276.56	0.41
	leng	19.91	1	19.91	10415.89	0.25
	dim:resp	10.43	9	1.16	606.25	0.15
	dim:leng	7.40	3	2.47	1289.44	0.11
	resp:leng	3.98	3	1.33	693.64	0.06

	dim:resp:leng	3.88	9	0.43	225.34	0.06
	residual	61.11	31968	0.00		

ANOVA Results for Disacquiescence Responding

Table C4

Guttman Errors ANOVA Table for Disacquiescence

	Parameter	Sum Sq	df	Mean Sq.	F value	Partial Omega Sq
Sensitivity	dim	24.30	3	8.10	7024.17	0.40
	resp	69.82	3	23.27	20185.53	0.65
	leng	0.06	1	0.06	56.26	0.00
	dim:resp	29.67	9	3.30	2859.27	0.45
	dim:leng	5.26	3	1.75	1520.65	0.12
	resp:leng	4.45	3	1.48	1285.43	0.11
	dim:resp:leng	5.35	9	0.59	515.70	0.13
	residual	36.86	31968	0.00		
Specificity	dim	0.25	3	0.08	16939.57	0.61
	resp	0.81	3	0.27	54891.78	0.84
	leng	0.02	1	0.02	4715.87	0.13
	dim:resp	0.35	9	0.04	8020.13	0.69
	dim:leng	0.06	3	0.02	4397.32	0.29
	resp:leng	0.05	3	0.02	3195.54	0.23
	dim:resp:leng	0.15	9	0.02	3481.46	0.49
	residual	0.16	31968	0.00		
NPV	dim	0.27	3	0.09	8315.81	0.44
	resp	0.77	3	0.26	24239.66	0.69
	leng	0.00	1	0.00	12.89	0.00
	dim:resp	0.32	9	0.04	3374.51	0.49
	dim:leng	0.06	3	0.02	1838.13	0.15
	resp:leng	0.05	3	0.02	1527.24	0.13
	dim:resp:leng	0.06	9	0.01	632.31	0.15
	residual	0.34	31968	0.00		
PPV	dim	89.77	3	29.92	13715.21	0.56
	resp	248.24	3	82.75	37926.63	0.78
	leng	0.14	1	0.14	62.73	0.00
	dim:resp	103.15	9	11.46	5253.38	0.60
	dim:leng	16.45	3	5.48	2512.82	0.19

	resp:leng	16.05	3	5.35	2452.57	0.19
	dim:resp:leng	20.18	9	2.24	1027.61	0.22
	residual	69.75	31968	0.00		

Table C5

H^T_i ANOVA Table for Disacquiescence

	Parameter	Sum Sq	df	Mean Sq.	F value	Partial Omega Sq
Sensitivity	dim	0.97	3	0.32	816.87	0.07
	resp	20.19	3	6.73	16947.98	0.61
	leng	1.75	1	1.75	4412.91	0.12
	dim:resp	0.61	9	0.07	171.55	0.05
	dim:leng	2.08	3	0.69	1743.32	0.14
	resp:leng	2.08	3	0.69	1749.93	0.14
	dim:resp:leng	4.30	9	0.48	1202.45	0.25
	residual	12.69	31968	0.00		
Specificity	dim	0.01	3	0.00	182.67	0.02
	resp	0.27	3	0.09	3981.32	0.27
	leng	0.02	1	0.02	795.91	0.02
	dim:resp	0.02	9	0.00	75.94	0.02
	dim:leng	0.03	3	0.01	513.11	0.05
	resp:leng	0.03	3	0.01	409.34	0.04
	dim:resp:leng	0.05	9	0.01	243.38	0.06
	residual	0.73	31968	0.00		
NPV	dim	0.01	3	0.00	938.71	0.08
	resp	0.23	3	0.08	19618.20	0.65
	leng	0.02	1	0.02	4998.42	0.14
	dim:resp	0.01	9	0.00	205.94	0.05
	dim:leng	0.02	3	0.01	2070.11	0.16
	resp:leng	0.02	3	0.01	1993.40	0.16
	dim:resp:leng	0.05	9	0.01	1359.42	0.28
	residual	0.12	31968	0.00		
PPV	dim	3.77	3	1.26	1021.52	0.09
	resp	76.87	3	25.62	20846.43	0.66
	leng	6.36	1	6.36	5173.77	0.14
	dim:resp	2.31	9	0.26	209.24	0.06

	dim:leng	8.03	3	2.68	2178.77	0.17
	resp:leng	7.75	3	2.58	2101.41	0.16
	dim:resp:leng	16.36	9	1.82	1478.63	0.29
	residual	39.30	31968	0.00		

Table C6

U3 ANOVA Table for Disacquiescence

	Parameter	**Sum Sq**	**df**	**Mean Sq**	**F value**	**Partial Omega Sq**
Sensitivity	dim	4.40	3	1.47	1342.90	0.11
	resp	32.77	3	10.92	10003.42	0.48
	leng	1.44	1	1.44	1320.64	0.04
	dim:resp	20.33	9	2.26	2068.61	0.37
	dim:leng	9.58	3	3.19	2922.99	0.22
	resp:leng	15.51	3	5.17	4734.25	0.31
	dim:resp:leng	6.14	9	0.68	624.78	0.15
	residual	34.91	31968	0.00		
Specificity	dim	0.04	3	0.01	1632.23	0.13
	resp	0.41	3	0.14	15157.74	0.59
	leng	0.02	1	0.02	1705.38	0.05
	dim:resp	0.26	9	0.03	3140.64	0.47
	dim:leng	0.12	3	0.04	4240.08	0.28
	resp:leng	0.19	3	0.06	6830.48	0.39
	dim:resp:leng	0.08	9	0.01	921.31	0.21
	residual	0.29	31968	0.00		
NPV	dim	0.05	3	0.02	1595.44	0.13
	resp	0.37	3	0.12	12136.10	0.53
	leng	0.02	1	0.02	1580.05	0.05
	dim:resp	0.23	9	0.03	2510.50	0.41
	dim:leng	0.11	3	0.04	3525.92	0.25
	resp:leng	0.17	3	0.06	5722.31	0.35
	dim:resp:leng	0.07	9	0.01	755.78	0.18
	residual	0.32	31968	0.00		
PPV	dim	16.49	3	5.50	2349.11	0.18
	resp	125.81	3	41.94	17918.36	0.63
	leng	5.57	1	5.57	2379.45	0.07
	dim:resp	77.59	9	8.62	3683.55	0.51

dim:leng	37.22	3	12.41	5300.20	0.33
resp:leng	58.57	3	19.52	8341.53	0.44
dim:resp:leng	23.09	9	2.57	1095.95	0.24
residual	74.82	31968	0.00		

ANOVA Results for Midpoint Responding

Table C7

Guttman Errors ANOVA Table for Midpoint Responding

	Parameter	Sum Sq	df	Mean Sq	F value	Partial Omega Sq
Sensitivity	dim	0.00	3	0.00	84.70	0.01
	resp	0.03	3	0.01	952.26	0.08
	leng	0.00	1	0.00	258.31	0.01
	dim:resp	0.01	9	0.00	60.73	0.02
	dim:leng	0.00	3	0.00	15.57	0.00
	resp:leng	0.00	3	0.00	13.61	0.00
	dim:resp:leng	0.00	9	0.00	37.96	0.01
	residual	0.33	31968	0.00		
Specificity	dim	0.01	3	0.00	992.26	0.09
	resp	0.02	3	0.01	1189.62	0.10
	leng	0.00	1	0.00	830.30	0.03
	dim:resp	0.03	9	0.00	693.21	0.16
	dim:leng	0.01	3	0.00	695.10	0.06
	resp:leng	0.02	3	0.01	1315.30	0.11
	dim:resp:leng	0.03	9	0.00	697.07	0.16
	residual	0.16	31968	0.00		
NPV	dim	0.00	3	0.00	688.79	0.06
	resp	0.00	3	0.00	1792.49	0.14
	leng	0.00	1	0.00	20.55	0.00
	dim:resp	0.00	9	0.00	228.69	0.06
	dim:leng	0.00	3	0.00	288.19	0.03
	resp:leng	0.00	3	0.00	571.66	0.05
	dim:resp:leng	0.00	9	0.00	304.31	0.08
	residual	0.00	31968	0.00		
PPV	dim	0.01	3	0.00	89.71	0.01
	resp	0.11	3	0.04	967.76	0.08
	leng	0.01	1	0.01	252.07	0.01
	dim:resp	0.02	9	0.00	61.99	0.02

	dim:leng	0.00	3	0.00	16.34	0.00
	resp:leng	0.00	3	0.00	13.53	0.00
	dim:resp:leng	0.01	9	0.00	40.65	0.01
	residual	1.20	31968	0.00		

Table C8

H^T_i ANOVA Table for Midpoint Responding

	Parameter	**Sum Sq**	**df**	**Mean Sq**	**F value**	**Partial Omega Sq**
Sensitivity	dim	48.07	3	16.02	16738.13	0.61
	resp	8.63	3	2.88	3004.80	0.22
	leng	19.81	1	19.81	20698.40	0.39
	dim:resp	32.84	9	3.65	3811.88	0.52
	dim:leng	15.56	3	5.19	5419.34	0.34
	resp:leng	5.44	3	1.81	1894.71	0.15
	dim:resp:leng	15.71	9	1.75	1823.74	0.34
	residual	30.60	31968	0.00		
Specificity	dim	0.60	3	0.20	10885.32	0.51
	resp	0.10	3	0.03	1787.01	0.14
	leng	0.23	1	0.23	12735.90	0.28
	dim:resp	0.40	9	0.04	2437.63	0.41
	dim:leng	0.20	3	0.07	3637.32	0.25
	resp:leng	0.06	3	0.02	1028.24	0.09
	dim:resp:leng	0.21	9	0.02	1297.45	0.27
	residual	0.59	31968	0.00		
NPV	dim	0.53	3	0.18	19105.86	0.64
	resp	0.10	3	0.03	3395.67	0.24
	leng	0.22	1	0.22	23512.43	0.42
	dim:resp	0.36	9	0.04	4337.59	0.55
	dim:leng	0.17	3	0.06	6201.03	0.37
	resp:leng	0.06	3	0.02	2123.98	0.17
	dim:resp:leng	0.18	9	0.02	2096.11	0.37
	residual	0.30	31968	0.00		
PPV	dim	182.00	3	60.67	24931.57	0.70
	resp	33.13	3	11.04	4538.74	0.30
	leng	73.62	1	73.62	30254.00	0.49
	dim:resp	127.20	9	14.13	5808.20	0.62
	dim:leng	60.20	3	20.07	8245.94	0.44
	resp:leng	20.67	3	6.89	2830.86	0.21

	dim:resp:leng	62.43	9	6.94	2850.50	0.44
	residual	77.79	31968	0.00		

Table C9

U3 ANOVA Table for Midpoint Responding

	Parameter	Sum Sq	df	Mean Sq	F value	Partial Omega Sq
Sensitivity	dim	0.01	3	0.00	210.47	0.02
	resp	0.06	3	0.02	1183.74	0.10
	leng	0.00	1	0.00	119.11	0.00
	dim:resp	0.08	9	0.01	508.57	0.12
	dim:leng	0.00	3	0.00	26.98	0.00
	resp:leng	0.00	3	0.00	29.97	0.00
	dim:resp:leng	0.00	9	0.00	25.34	0.01
	residual	0.56	31968	0.00		
Specificity	dim	0.00	3	0.00	60.62	0.01
	resp	0.00	3	0.00	112.91	0.01
	leng	0.00	1	0.00	10.93	0.00
	dim:resp	0.01	9	0.00	108.26	0.03
	dim:leng	0.00	3	0.00	67.01	0.01
	resp:leng	0.00	3	0.00	85.92	0.01
	dim:resp:leng	0.01	9	0.00	69.11	0.02
	residual	0.39	31968	0.00		
NPV	dim	0.00	3	0.00	259.86	0.02
	resp	0.00	3	0.00	1001.99	0.09
	leng	0.00	1	0.00	32.49	0.00
	dim:resp	0.00	9	0.00	450.09	0.11
	dim:leng	0.00	3	0.00	61.42	0.01
	resp:leng	0.00	3	0.00	57.12	0.01
	dim:resp:leng	0.00	9	0.00	55.90	0.02
	residual	0.01	31968	0.00		
PPV	dim	0.04	3	0.01	221.21	0.02
	resp	0.23	3	0.08	1183.69	0.10
	leng	0.01	1	0.01	117.38	0.00
	dim:resp	0.31	9	0.03	535.86	0.13
	dim:leng	0.01	3	0.00	28.92	0.00
	resp:leng	0.01	3	0.00	26.92	0.00

	dim:resp:leng	0.02	9	0.00	26.31	0.01
	residual	2.08	31968	0.00		

ANOVA Results for Extreme Responding

Table C10

Guttman Errors ANOVA Table for Extreme Responding

	Parameter	Sum Sq	df	Mean Sq	F value	Partial Omega Sq
Sensitivity	dim	3.29	3	1.10	573.45	0.05
	resp	123.16	3	41.05	21477.94	0.67
	leng	86.32	1	86.32	45160.06	0.59
	dim:resp	6.66	9	0.74	386.96	0.10
	dim:leng	6.44	3	2.15	1123.43	0.10
	resp:leng	4.79	3	1.60	834.76	0.07
	dim:resp:leng	6.30	9	0.70	366.39	0.09
	residual	61.11	31968	0.00		
Specificity	dim	0.03	3	0.01	1885.75	0.15
	resp	1.52	3	0.51	95468.71	0.90
	leng	1.17	1	1.17	220747.25	0.87
	dim:resp	0.07	9	0.01	1565.09	0.31
	dim:leng	0.09	3	0.03	5915.83	0.36
	resp:leng	0.07	3	0.02	4430.32	0.29
	dim:resp:leng	0.06	9	0.01	1344.61	0.27
	residual	0.17	31968	0.00		
NPV	dim	0.04	3	0.01	654.20	0.06
	resp	1.37	3	0.46	25060.03	0.70
	leng	0.97	1	0.97	53085.50	0.62
	dim:resp	0.07	9	0.01	445.70	0.11
	dim:leng	0.07	3	0.02	1329.45	0.11
	resp:leng	0.05	3	0.02	990.22	0.08
	dim:resp:leng	0.07	9	0.01	420.32	0.11
	residual	0.58	31968	0.00		
PPV	dim	10.79	3	3.60	1909.56	0.15
	resp	471.07	3	157.02	83334.34	0.89
	leng	348.04	1	348.04	184708.93	0.85
	dim:resp	24.60	9	2.73	1450.43	0.29
	dim:leng	26.73	3	8.91	4729.54	0.31
	resp:leng	19.07	3	6.36	3372.75	0.24

	dim:resp:leng	21.59	9	2.40	1272.95	0.26
	residual	60.24	31968	0.00		

Table C11

H^T_i ANOVA Table for Extreme Responding

	Parameter	**Sum Sq**	**df**	**Mean Sq**	**F value**	**Partial Omega Sq**
Sensitivity	dim	0.43	3	0.14	298.19	0.03
	resp	5.48	3	1.83	3756.37	0.26
	leng	2.51	1	2.51	5154.08	0.14
	dim:resp	1.08	9	0.12	247.21	0.06
	dim:leng	1.77	3	0.59	1210.87	0.10
	resp:leng	0.71	3	0.24	488.70	0.04
	dim:resp:leng	1.20	9	0.13	275.04	0.07
	residual	15.54	31968	0.00		
Specificity	dim	0.01	3	0.00	51.43	0.00
	resp	0.07	3	0.02	538.55	0.05
	leng	0.02	1	0.02	523.01	0.02
	dim:resp	0.02	9	0.00	49.67	0.01
	dim:leng	0.02	3	0.01	197.88	0.02
	resp:leng	0.02	3	0.01	131.43	0.01
	dim:resp:leng	0.01	9	0.00	40.12	0.01
	residual	1.31	31968	0.00		
NPV	dim	0.00	3	0.00	347.16	0.03
	resp	0.06	3	0.02	4316.06	0.29
	leng	0.03	1	0.03	5743.01	0.15
	dim:resp	0.01	9	0.00	286.61	0.07
	dim:leng	0.02	3	0.01	1410.19	0.12
	resp:leng	0.01	3	0.00	603.57	0.05
	dim:resp:leng	0.01	9	0.00	307.44	0.08
	residual	0.15	31968	0.00		
PPV	dim	1.55	3	0.52	328.26	0.03
	resp	21.12	3	7.04	4464.83	0.30
	leng	9.07	1	9.07	5750.83	0.15
	dim:resp	4.29	9	0.48	302.58	0.08
	dim:leng	6.64	3	2.21	1404.51	0.12
	resp:leng	2.91	3	0.97	614.50	0.05
	dim:resp:leng	4.57	9	0.51	322.15	0.08
	residual	50.40	31968	0.00		

Table C12

U3 ANOVA Table for Extreme Responding

	Parameter	Sum Sq	df	Mean Sq	F value	Partial Omega Sq
Sensitivity	dim	2.42	3	0.81	298.52	0.03
	resp	175.19	3	58.40	21620.95	0.67
	leng	105.39	1	105.39	39020.97	0.55
	dim:resp	13.54	9	1.50	556.88	0.14
	dim:leng	1.47	3	0.49	181.94	0.02
	resp:leng	8.84	3	2.95	1091.15	0.09
	dim:resp:leng	5.33	9	0.59	219.30	0.06
	residual	86.34	31968	0.00		
Specificity	dim	0.02	3	0.01	292.28	0.03
	resp	1.96	3	0.65	31156.24	0.74
	leng	1.26	1	1.26	60029.68	0.65
	dim:resp	0.18	9	0.02	931.20	0.21
	dim:leng	0.01	3	0.00	185.32	0.02
	resp:leng	0.11	3	0.04	1749.54	0.14
	dim:resp:leng	0.07	9	0.01	351.83	0.09
	residual	0.67	31968	0.00		
NPV	dim	0.03	3	0.01	351.25	0.03
	resp	1.94	3	0.65	26198.99	0.71
	leng	1.17	1	1.17	47495.06	0.60
	dim:resp	0.15	9	0.02	685.43	0.16
	dim:leng	0.02	3	0.01	213.84	0.02
	resp:leng	0.10	3	0.03	1331.90	0.11
	dim:resp:leng	0.06	9	0.01	268.06	0.07
	residual	0.79	31968	0.00		
PPV	dim	8.38	3	2.79	1133.55	0.10
	resp	659.88	3	219.96	89282.49	0.89
	leng	407.49	1	407.49	165404.04	0.84
	dim:resp	52.95	9	5.88	2387.94	0.40
	dim:leng	4.99	3	1.66	674.98	0.06
	resp:leng	33.60	3	11.20	4546.30	0.30
	dim:resp:leng	20.33	9	2.26	916.83	0.20
	residual	78.76	31968	0.00		

ANOVA Results for Socially Desirable Responding

Table C13

Guttman Errors ANOVA Table for Socially Desirable Responding

	Parameter	Sum Sq	df	Mean Sq	F value	Partial Omega Sq
Sensitivity	dim	3.59	3	1.20	1704.97	0.14
	resp	5.63	3	1.88	2673.68	0.20
	leng	4.74	1	4.74	6757.33	0.17
	dim:resp	6.75	9	0.75	1068.90	0.23
	dim:leng	2.31	3	0.77	1097.90	0.09
	resp:leng	2.74	3	0.91	1300.91	0.11
	dim:resp:leng	7.11	9	0.79	1125.53	0.24
	residual	22.44	31968	0.00		
Specificity	dim	0.03	3	0.01	1887.35	0.15
	resp	0.08	3	0.03	4752.01	0.31
	leng	0.13	1	0.13	21977.76	0.41
	dim:resp	0.08	9	0.01	1565.33	0.31
	dim:leng	0.02	3	0.01	1085.60	0.09
	resp:leng	0.03	3	0.01	1900.22	0.15
	dim:resp:leng	0.16	9	0.02	3089.01	0.46
	residual	0.18	31968	0.00		
NPV	dim	0.04	3	0.01	2022.92	0.16
	resp	0.06	3	0.02	3314.09	0.24
	leng	0.06	1	0.06	9028.32	0.22
	dim:resp	0.07	9	0.01	1227.82	0.26
	dim:leng	0.02	3	0.01	1292.65	0.11
	resp:leng	0.03	3	0.01	1557.06	0.13
	dim:resp:leng	0.08	9	0.01	1417.33	0.28
	residual	0.20	31968	0.00		
PPV	dim	12.64	3	4.21	2302.21	0.18
	resp	20.76	3	6.92	3781.90	0.26
	leng	20.12	1	20.12	10996.51	0.26
	dim:resp	22.90	9	2.54	1390.74	0.28
	dim:leng	7.99	3	2.66	1455.03	0.12
	resp:leng	10.92	3	3.64	1989.41	0.16
	dim:resp:leng	28.50	9	3.17	1731.06	0.33
	residual	58.48	31968	0.00		

Table C14

H^T_i ANOVA Table for Socially Desirable Responding

	Parameter	Sum Sq	df	Mean Sq	F value	Partial Omega Sq
Sensitivity	dim	3.47	3	1.16	1572.88	0.13
	resp	69.73	3	23.24	31640.67	0.75
	leng	9.24	1	9.24	12577.59	0.28
	dim:resp	10.26	9	1.14	1551.74	0.30
	dim:leng	9.43	3	3.14	4280.95	0.29
	resp:leng	15.31	3	5.10	6948.45	0.39
	dim:resp:leng	6.16	9	0.68	932.01	0.21
	residuals	23.48	31968	0.00		
Specificity	dim	0.04	3	0.01	486.39	0.04
	resp	0.85	3	0.28	10203.90	0.49
	leng	0.09	1	0.09	3378.78	0.10
	dim:resp	0.18	9	0.02	716.06	0.17
	dim:leng	0.16	3	0.05	1879.71	0.15
	resp:leng	0.18	3	0.06	2135.72	0.17
	dim:resp:leng	0.10	9	0.01	396.00	0.10
	residuals	0.89	31968	0.00		
NPV	dim	0.04	3	0.01	1863.92	0.15
	resp	0.77	3	0.26	37871.21	0.78
	leng	0.10	1	0.10	14832.02	0.32
	dim:resp	0.12	9	0.01	1922.80	0.35
	dim:leng	0.11	3	0.04	5286.34	0.33
	resp:leng	0.17	3	0.06	8288.79	0.44
	dim:resp:leng	0.07	9	0.01	1135.73	0.24
	residuals	0.22	31968	0.00		
PPV	dim	11.35	3	3.78	1734.51	0.14
	resp	280.04	3	93.35	42805.84	0.80
	leng	31.45	1	31.45	14422.55	0.31
	dim:resp	48.89	9	5.43	2491.05	0.41
	dim:leng	41.31	3	13.77	6314.95	0.37
	resp:leng	53.91	3	17.97	8241.15	0.44
	dim:resp:leng	28.24	9	3.14	1439.11	0.29
	residuals	69.71	31968	0.00		

Table C15

U3 ANOVA Table for Socially Desirable Responding

	Parameter	Sum Sq	df	Mean Sq	F value	Partial Omega Sq
Sensitivity	dim	2.98	3	0.99	1376.70	0.11
	resp	6.37	3	2.12	2939.20	0.22
	leng	2.11	1	2.11	2927.80	0.08
	dim:resp	7.67	9	0.85	1180.49	0.25
	dim:leng	0.57	3	0.19	263.62	0.02
	resp:leng	1.59	3	0.53	731.77	0.06
	dim:resp:leng	3.03	9	0.34	466.20	0.12
	residual	23.08	31968	0.00		
Specificity	dim	0.04	3	0.01	276.89	0.03
	resp	0.08	3	0.03	564.49	0.05
	leng	0.01	1	0.01	272.20	0.01
	dim:resp	0.10	9	0.01	240.33	0.06
	dim:leng	0.02	3	0.01	142.05	0.01
	resp:leng	0.03	3	0.01	183.28	0.02
	dim:resp:leng	0.04	9	0.00	105.53	0.03
	residual	1.50	31968	0.00		
NPV	dim	0.03	3	0.01	1927.30	0.15
	resp	0.07	3	0.02	4108.04	0.28
	leng	0.02	1	0.02	3857.70	0.11
	dim:resp	0.09	9	0.01	1665.75	0.32
	dim:leng	0.01	3	0.00	411.49	0.04
	resp:leng	0.02	3	0.01	1044.89	0.09
	dim:resp:leng	0.03	9	0.00	652.11	0.15
	residual	0.18	31968	0.00		
PPV	dim	12.08	3	4.03	2334.66	0.18
	resp	23.61	3	7.87	4564.38	0.30
	leng	7.47	1	7.47	4333.48	0.12
	dim:resp	28.30	9	3.14	1823.82	0.34
	dim:leng	2.56	3	0.85	495.29	0.04
	resp:leng	6.36	3	2.12	1229.31	0.10
	dim:resp:leng	11.78	9	1.31	758.88	0.18
	residual	55.12	31968	0.00		

ANOVA Results for Careless Responding

Table C16

Guttman Errors ANOVA Table for Careless Responding

	Parameter	Sum Sq	df	Mean Sq	F value	Partial Omega Sq
Sensitivity	dim	17.84	3	5.95	2960.94	0.22
	resp	644.49	3	214.83	106939.81	0.91
	leng	81.71	1	81.71	40674.35	0.56
	dim:resp	31.83	9	3.54	1760.33	0.33
	dim:leng	5.26	3	1.75	872.56	0.08
	resp:leng	13.75	3	4.58	2281.59	0.18
	dim:resp:leng	13.43	9	1.49	742.94	0.17
	residual	64.22	31968	0.00		
Specificity	dim	0.22	3	0.07	19315.19	0.64
	resp	7.96	3	2.65	688161.26	0.98
	leng	1.01	1	1.01	262334.62	0.89
	dim:resp	0.41	9	0.05	11823.14	0.77
	dim:leng	0.07	3	0.02	6029.79	0.36
	resp:leng	0.16	3	0.05	13431.94	0.56
	dim:resp:leng	0.16	9	0.02	4708.92	0.57
	residual	0.12	31968	0.00		
NPV	dim	0.20	3	0.07	3403.42	0.24
	resp	7.16	3	2.39	122954.56	0.92
	leng	0.91	1	0.91	46777.46	0.59
	dim:resp	0.35	9	0.04	2030.76	0.36
	dim:leng	0.06	3	0.02	1008.85	0.09
	resp:leng	0.15	3	0.05	2604.00	0.20
	dim:resp:leng	0.15	9	0.02	852.86	0.19
	residual	0.62	31968	0.00		
PPV	dim	72.46	3	24.15	21504.47	0.67
	resp	2500.70	3	833.57	742107.89	0.99
	leng	316.36	1	316.36	281651.88	0.90
	dim:resp	126.81	9	14.09	12543.76	0.78
	dim:leng	21.32	3	7.11	6326.89	0.37
	resp:leng	51.01	3	17.00	15138.66	0.59
	dim:resp:leng	51.57	9	5.73	5101.34	0.59
	residual	35.91	31968	0.00		

Table C17

H^T_i ANOVA Table for Careless Responding

	Parameter	Sum Sq	df	Mean Sq	F value	Partial Omega Sq
Sensitivity	dim	79.82	3	26.61	28848.72	0.73
	resp	124.74	3	41.58	45082.90	0.81
	leng	25.88	1	25.88	28063.65	0.47
	dim:resp	19.95	9	2.22	2403.37	0.40
	dim:leng	4.07	3	1.36	1472.67	0.12
	resp:leng	17.72	3	5.91	6402.67	0.38
	dim:resp:leng	6.27	9	0.70	755.08	0.17
	residual	29.48	31968	0.00		
Specificity	dim	0.94	3	0.31	10524.20	0.50
	resp	1.59	3	0.53	17764.28	0.62
	leng	0.32	1	0.32	10885.75	0.25
	dim:resp	0.24	9	0.03	898.70	0.20
	dim:leng	0.05	3	0.02	541.72	0.05
	resp:leng	0.19	3	0.06	2140.67	0.17
	dim:resp:leng	0.07	9	0.01	245.72	0.06
	residual	0.95	31968	0.00		
NPV	dim	0.88	3	0.29	32070.46	0.75
	resp	1.39	3	0.46	50503.30	0.83
	leng	0.29	1	0.29	31437.34	0.50
	dim:resp	0.22	9	0.02	2677.36	0.43
	dim:leng	0.05	3	0.02	1639.73	0.13
	resp:leng	0.19	3	0.06	7051.93	0.40
	dim:resp:leng	0.07	9	0.01	828.89	0.19
	residual	0.29	31968	0.00		
PPV	dim	303.60	3	101.20	46342.98	0.81
	resp	476.27	3	158.76	72701.60	0.87
	leng	96.10	1	96.10	44006.67	0.58
	dim:resp	76.28	9	8.48	3881.13	0.52
	dim:leng	14.98	3	4.99	2286.09	0.18
	resp:leng	67.63	3	22.54	10322.80	0.49
	dim:resp:leng	23.28	9	2.59	1184.64	0.25
	residual	69.81	31968	0.00		

Table C18

U3 ANOVA Table for Careless Responding

	Parameter	Sum Sq	df	Mean Sq	F value	Partial Omega Sq
Sensitivity	dim	13.09	3	4.36	2307.57	0.18
	resp	687.35	3	229.12	121161.49	0.92
	leng	89.64	1	89.64	47402.65	0.60
	dim:resp	23.96	9	2.66	1407.55	0.28
	dim:leng	8.60	3	2.87	1515.52	0.12
	resp:leng	39.44	3	13.15	6952.13	0.39
	dim:resp:leng	23.22	9	2.58	1364.33	0.28
	residual	60.45	31968	0.00		
Specificity	dim	0.15	3	0.05	3449.22	0.24
	resp	8.40	3	2.80	196326.58	0.95
	leng	1.12	1	1.12	78650.58	0.71
	dim:resp	0.30	9	0.03	2334.78	0.40
	dim:leng	0.10	3	0.03	2340.87	0.18
	resp:leng	0.47	3	0.16	10876.88	0.50
	dim:resp:leng	0.29	9	0.03	2229.66	0.39
	residual	0.46	31968	0.00		
NPV	dim	0.14	3	0.05	2635.30	0.20
	resp	7.63	3	2.54	139624.16	0.93
	leng	1.00	1	1.00	54739.57	0.63
	dim:resp	0.27	9	0.03	1626.71	0.31
	dim:leng	0.10	3	0.03	1738.77	0.14
	resp:leng	0.44	3	0.15	7982.67	0.43
	dim:resp:leng	0.26	9	0.03	1572.13	0.31
	residual	0.58	31968	0.00		
PPV	dim	51.32	3	17.11	12129.69	0.53
	resp	2661.46	3	887.15	629064.20	0.98
	leng	347.28	1	347.28	246249.80	0.88
	dim:resp	90.50	9	10.06	7130.00	0.67
	dim:leng	33.15	3	11.05	7835.09	0.42
	resp:leng	150.54	3	50.18	35580.68	0.77
	dim:resp:leng	89.09	9	9.90	7018.91	0.66
	residual	45.08	31968	0.00		