ISSN 1742-5867

Stammering Research

Editor Peter Howell PhD FIOA
Professor of Experimental Psychology
University College London

The British Stammering Association

The British Stammering Association
15 Old Ford Road
London E2 9PJ

Tel: 020 8983 1003
Fax: 020 8983 3591
Helpline: 0845 603 2001

www.stammering.org

Reg Charity No 1089967
Reg Company No 4297778

Stammering Research

A Journal Published by the British Stammering Association

Stammering Research

A Journal Published by the British Stammering Association

Volume 1, Issue 4, January 2005

Contents

Notice

The British Stammering Association is a UK-based charity which seeks to promote understanding into the causes, treatment and understanding of stammering. Its activities include research into stammering which it supports through its vacation studentship scheme (http://www.stammering.org/research_schol.html) and the publication of Stammering Research (provided free of charge to all-comers).

Stammering Research is intended to promote public understanding of high quality scientific research into stammering and allied areas

If individuals wish to make a donation to support either of these initiatives, they should forward a cheque (payable to the British Stammering Association) to The British Stammering Association, 15 Old Ford Road, London E2 9PJ, or call the BSA on 020 8983 1003 (+44 20 8983 1003 from abroad) with their credit card details. If they wish this to be used specifically for either the vacation studentship scheme or Stammering Research, they should mark it accordingly on the back of the cheque. For information on tax-effective ways to support the charity's research activities, please go to http://www.stammering.org/donations.html.

Donors will be listed in the last issue of the appropriate volume of the journal unless they indicate otherwise. Companies wishing to make a donation or who wish to make enquiries about advertising in Stammering Research should address correspondence to Norbert Lieckfeldt at nl@stammering.org.

‘Stammering Research’.
An on-line journal published by the British Stammering Association
ISSN 1742-5867

Description

Stammering Research is an international journal published in electronic format. Currently it appears as four quarterly issues per volume (officially published March 31st, June 30th, September 30th and December 31st). The first issue of volume one appeared on March 31st 2004. The journal is dedicated to the furtherance of research into stammering, and is published under the auspices of the British Stammering Association. It seeks reports of significant pieces of work on stammering and allied areas, such as other speech disorders and disfluency in the spontaneous speech of fluent speakers. Articles published include (though are not be limited to) reviews in an area in which the author has produced eminent work and attempts to introduce new techniques into studies in the field. The journal offers an opportunity to table topics where there are grounds for considering a major rethink is required, as well as detailing development and assessment of research-based techniques for diagnosis and treatment of the disorder. Submissions are encouraged that facilitate open access to scientific materials and tools. Articles are peer-reviewed, the role of reviewers being to ensure that accepted standards of scientific reporting are met, including correction of factual errors. Disagreements about interpretation of findings raised by reviewers will be passed on by the editorial board to the authors of accepted papers. These disagreements will not necessarily preclude publication of the article if they are judged to be topics that are suitable for open peer commentary. Once accepted, commentaries will be sought (actively and by self-nomination) from specialists within the field of communication disorder and its allied disciplines. These commentaries will be reviewed for style and content. The author’s responses will be reviewed in the same way. The article, open peer commentaries and author’s responses will be published in the same issue as the target article and in subsequent issues. Authors should contact the editor in the first instance with a short description of the topic area so that its general suitability can be assessed before full submission. Notification that a topic is suitable does not imply that the paper that is subsequently submitted will be accepted. Decisions about suitability will be made by the editorial board.

Editor
Peter Howell
Department of Psychology
University College London
Gower St.,
London WC1E 6BT
England

SUMMARY OF STEP BY STEP PROCEDURE FOR AUTHORING AN ARTICLE TO STAMMERING RESEARCH

1. Contact the editor with a brief outline of the proposed article. The editor and other board members make initial decisions only as to the suitability of the general area proposed. The primary function in this step is to ensure the topic is of sufficiently broad interest for, and within the remit of, the readership of Stammering Research. The intent behind this initial contact is to ensure authors do not spend time preparing articles on unsuitable topics. Review, empirical and theoretical work are all appropriate. Authors will be informed whether the judgement is that the proposed topic has a suitable, or too narrow, a focus. Indication that the scope is too narrow does not imply anything about the scientific standard of the proposed work. Neither does notification that a topic is suitable indicate that the submitted work will necessarily be accepted for publication (all submitted material has to go through the normal processes of peer review). Articles should be submitted by email to p.howell@ucl.ac.uk.
2. Submitted articles are peer reviewed in the normal way and an indication as to suitability of publication or not (possibly after revision) is notified to the author by the editor.
3. After an article has been accepted, the author cannot change the article. It is then made available for open peer commentary. Details how the accepted article can be accessed are posted on the British Stammering Association's website (www.stammering.org). Indications that the article is available for access are posted on http://www.mnsu.edu/dept/comdis/kuster/Internet/Listserv.html for ASHA members, the British Stammering Association's website (http://www.stammering.org, the stut-l list (stutt-l@listmail. temple.edu), the stutt-x mailing list (stutt-x@asu.edu), and on the stuttering home page (www.stutteringhomepage.com). The primary function in posting details about access available to an accepted article, is to alert potential commentators. A list of commentators is being drawn up and individuals are encouraged to submit their nominations (for themselves or others).
4. See the next page for precise details how to prepare a commentary and the timetable allowed for this. When preparing a commentary, authors might find it helpful to consult a recent issue of Stammering Research to see the range of comments that are appropriate, the style and format of commentary
5. All accepted commentaries are available to the author of a target article from receipt until two weeks after invitations for commentaries has closed. In this time, the author can prepare a response to commentaries. The response will be peer-reviewed by the editorial board. Further details are given on the next page and authors should again consult a recent issue of Stammering Research to see the sorts of comments that are appropriate, style and formatting of a submission.
6. On completion of this process, the target article, commentaries and response to commentaries will be published together in Stammering Research. Authors are responsible for preparing their articles according to the stipulated format. The current and previous issues of the journal are available as PDF files at http://www.speech.psychol.ucl.ac.uk/.

Notes about commentaries for Stammering Research
ISSN 1742-5867

Once a manuscript has been accepted as a target article, the authors cannot change it. The manuscript needs to be available for commentary before it is officially published so that commentaries and the author's responses can appear simultaneously.

Manuscripts are posted for commentary on http://www.psychol.ucl.ac.uk/ under *Stammering Research*. Commentators are alerted as indicated on the previous page.

Manuscripts will be available for peer commentary for six weeks. Commentaries have to reach the editor, or associate editor, responsible for the article within that time (late submissions will not be accepted). Commentaries should ordinarily not exceed a total (including references and other material) of 1,000 words. The commentaries have to conform to APA style conventions.

In order to appear in the same issue as the target article, commentaries should be sent by email as soon as possible within the period the article is open for peer commentary. The commentary should appear within the body of the email text (not as an attachment) and be sent to p.howell@ucl.ac.uk. Authors of target articles will receive commentaries as they are accepted and have two weeks from close of submission of commentaries to complete their responses. Commentaries that appear outside this timetable may appear as continuing commentaries in subsequent issues (these are considered in the same way as commentaries that appear at the same time as the target article.

Commentaries will be peer-reviewed and edited for style as well as content. Authors of commentaries need to establish the relevance of their submission to the target article at the outset, and preferably also show an awareness of the wider work of the target article's author.

If there are several commentaries which raise the same point, the editorial board reserves the right to group them together and prepare them as a single coauthored commentary. In this (probably rare) eventuality, the authors will have the opportunity to see the manuscript and decide whether they wish to be included on the list of authors.

Editing and revision of commentaries will be completed within two weeks of close of submission. Revisions that are not satisfactorily completed in this period, or that are received late, may be published as continuing commentaries.

Formatting Accepted Publications in Stammering Research

Peter Howell[1], John Smith[2], and John Doe[3]

[1]*Department of Psychology, University College London, Gower St., London WC1E 6BT England*
P.Howell@ucl.ac.uk

[2]*Stuttering Treatment Clinic, Somewhere, Some Country*
Smith@some.email.address

[3]*For private citizens, house number and street, city and postcode/ZIPcode, Country*
Doe@email.address.if.you.have.one

Abstract. A short abstract summarizing the significant content and contribution of the paper should be included here. This page illustrates and describes the format for paper submissions. Authors are requested to adhere as closely as possible to this format once an article is accepted. The abstract should be in Times New Roman 9-point font, justified with left and right margins indented 1 cm in from the margins of the main text. **Keywords**. A few key terms that will be used by abstracting services to make sure your article reaches those who will be interested in reading what you say.

1. Introduction

Articles and commentaries should be submitted for review in APA format. After an article or commentary is accepted, it needs to be prepared according to the journal format as indicated next. Articles and commentaries must be in Word format. An article will typically be up to **15,000 words**. A commentary should preferably be up to **1,000 words**. Authors may submit longer articles or commentaries for consideration but these may be reduced in length by the editor. Articles with fewer than 15,000 words and commentaries with fewer than 1,000 words are acceptable if the author can demonstrate sufficient content and contribution. Typically commentaries will have an abstract, usually only a single section in the text headed so as to identify the target article. If an author needs to use more than one section heading and diagrams or figures, then they should follow the same instructions as for preparation of a target article. Each page of an article should consist of single column, of single-spaced text in a 16cm x 24cm column using **A4** or **US Letter** settings on your word processor as illustrated in Figures 1 and 2. Figures should be numbered consecutively and appear close to the text where they are mentioned.

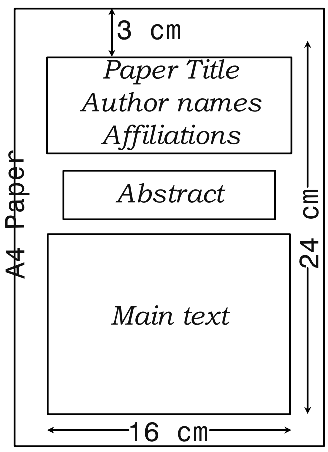


Figure 1: First page format

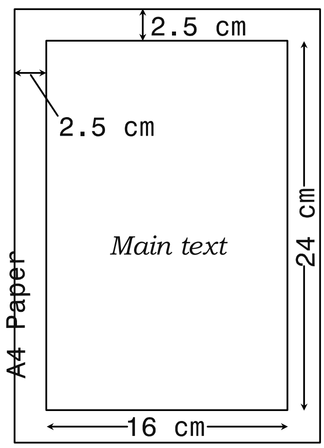


Figure 2: Subsequent page format

2. Detail of styles

The article or commentary title should be bold and centred using 14 point Times New Roman font. Authors' names, affiliations and email details should be centred using 10 point Times New Roman font. The author's name and affiliation should be italicized. The main text and the bibliographical references must be left and right justified and single line spaced. The main text should be in 10 point Times New Roman font with numbered section headings in 11 point bold font.

All references should be cited using APA referencing styles. For example a publication which is referred to as support for a statement would be cited in the text this way (Howell & Sackin, 2002) whatever the number of authors. When an article is referred to directly in the text as in "… in the work of Howell and Sackin (2002) the …" only the year is placed in brackets. If there is more than one reference from the same authors in the same year then they are distinguished by using different letter designations after the year as in 1996a, 1996b etc. In the references below, examples are given of how a conference paper, a journal paper and a book would be listed. All references should be listed at the end of the paper using 9 point Times New Roman font.

All figures, and diagrams must be good quality black and white images suitable for readers to display and print. Colour illustrations or text can be used, but bear in mind readers who want to print articles may not have access to a colour printer. When an article is accepted, figures and pictures must be inserted in the word file in the exact position they will appear in the publication. Any format for figures, pictures and diagrams may be used provided they allow good quality reproduction for readers who wish to print off a copy.

References

Howell, P. (2002). The EXPLAN theory of fluency control applied to the treatment of stuttering by altered feedback and operant procedures. In E. Fava (Ed.), *Pathology and therapy of speech disorders (pp. 95 - 118). Amsterdam: John Benjamins.*

Howell, P., & Sackin, S. (2002). Timing interference to speech in altered listening conditions. *Journal of the Acoustical Society of America*, **111**, 2842-2852.

Rosen, S., & Howell, P. (1991). *Signals and Systems for Speech and Hearing*. London and San Diego: Academic Press.

Stammering Research

A Journal Published by the British Stammering Association

Volume 1, Issue 4, January 2005

Contents

Editorial for Stammering Research

This is the fourth and final issue of volume 1 of *Stammering Research*. It includes two target articles and a report of an analysis of some of the Howell and Huckvale (2004) data that were published previously in the journal. The two target articles both concern statistical issues relevant to research into stammering. The first article, by Adrian Davis and Peter Howell, is an introduction to some basic statistical issues in the area and sets the scene for future statistical articles. The second article gives background details about one statistical approach (Structural Equation Modeling, SEM) that can be used to assess alternative multivariate models of various aspects of stammering. Statements to the effect that stammering is a disorder that is determined by a multiplicity of factors abound. There are few, if any, models that specify the structural relations between the various factors in a way that they can be evaluated against alternative models. Simply acknowledging that many variables affect stammering behavior (as in the Demands and Capacities model) does not indicate whether the variables have a significant effect, how each variable interacts with others that the author would wish to include and so on. SEM is one technique that allows such models, once constructed to be evaluated.

Both these articles are published as target articles and, as such, commentaries on them will be reviewed on receipt. This allows me to air, once more, some changes the editorial board has made to submission procedures for articles and commentaries. Submissions should now be sent by email to p.howell@ucl.ac.uk (this applies to prospective target articles as well as commentaries) and commentaries on any article will be received, reviewed and, if accepted, published in issues other than the one in which the target article appeared. Formerly commentaries appeared in the same issue as the target article but from feedback I received, it appeared that some commentators would have appreciated more time. Personally, I think it is important for authors to see immediately the impact a target article has had. For this reason, I would urge individuals to submit commentaries promptly. A second issue concerns what type of commentary would be appropriate for a statistical tutorial? The introductory article by Davis and Howell probably will not attract commentaries because of the nature of the material. The SEM paper probably will raise comments that need addressing. To pick two important ones, first, SEM presupposes a step of theory construction. Hopefully, we will begin to see precisely specified theories (maybe with statistical evaluation) as submissions to Stammering Research. It would be particularly nice to see models of treatment outcome. Second, the approach put forward in the SEM article is only one approach to these issues. Again, following the philosophy behind *Stammering Research*, we would like to see authors presenting alternative approaches (as potential target articles if they are substantive enough) so that a comprehensive box of tools can be built up for investigations in this area. A more general point *a propos* publication of these articles, is that we hope to see *Stammering Research* as the forum for introducing, tutoring and discussing a wide range of statistical issues that are pertinent to research in this area.

In volume 1, issue 2 of *Stammering Research*, a data archive was released of spontaneous monologue speech from speakers who stutter covering a wide age-range in alternative formats appropriate for use with different free software packages (Howell & Huckvale, 2004). It is intended that this database (the UCL archive of stuttered speech, or UCLASS for short) be used for a range of different analyses. The Howell (2005) article that appears in this issue of *Stammering Research*, is provided as an example how to report one type of analysis that can be done with audio data. It is hoped that the international community will take up the prospects and new challenges that have been opened up by release of the UCLASS and other data (Howell, Davis, Bartrip & Wormald, 2004) through the pages of *Stammering Research*. A workshop conference has been arranged for June 27th 2005 in London which will include tutorials describing ways to deal with the UCLASS material described in Howell and Huckvale (2004) with talks by Rose and MacWhinney (on CHILDES) and Huckvale (on SFS). This has been timed to precede the Oxford Dysfluency Conference so people can attend both. Details of the conference can be found at http://www.speech.psychol.ucl.ac.uk/.

At the end of this first year involved in editing and publishing *Stammering Research*, I feel there have been significant advances in research opportunities available to the whole spectrum of the journal's readers. The open peer commentary format is unique in this area of research and this journal provides the only forum I know of which releases software and data specifically intended to enhance and enable people's ability to research into stammering. I thank all of those who have made this venture possible.

Peter Howell, December 2004

References

Howell, P. (2005). The effect of using time intervals of different length for judgements about stuttering. *Stammering Research*, **1**, 364-374.

Howell, P., Davis, S., Bartrip, J. & Wormald, L. (2004). Effectiveness of frequency shifted feedback at reducing disfluency for linguistically easy, and difficult, sections of speech (original audio recordings included). *Stammering Research*, **1**, 309-315.

Howell, P. & Huckvale, M. (2004) Facilities to assist people to research into stammered speech. *Stammering Research*, **1**, 130-242.

TARGET ARTICLE

Elements of statistical treatment of speech and hearing science data

Adrian Davis[1] and Peter Howell[2]
[1]*Director NHS Newborn Hearing Screening Programme*
MRC Hearing and Communication Group
Humanities Building Block A, Level 3
School of Education
University of Manchester
Oxford Road
Manchester M13 9PL
adrian@mrchear.man.ac.uk
[2] *Department of Psychology, Centre for Human Communications, Institute of Cognitive Neuroscience, and Institute of Movement Neuroscience, University College London, Gower St., London WC1E 6BT, England*
p.howell@ucl.ac.uk

Abstract: Many of the statistical issues involved in speech and hearing research are shared with other areas of medicine. This article is the first in a series intended to stimulate examination of research data in speech and hearing areas using a wide variety of techniques. This article specifically deals with two essential, but elementary, issues. The first is concerned with experimental design and choice of test data. The second, defines and explains statistical terms, concentrating particularly on the inference to the population mean from the sample mean.
Keywords: Statistics, experimental design, choice of data, inferential statistics.

1. Background

This article is the first in a series that will discuss statistical treatment of data from studies into communication and its disorders. Application areas include where data from people with speaking or hearing disorders are available and researchers want to visualize or quantify how performance of these individuals relate to fluent populations. The range of topics that can be covered is massive starting with simple ones like how to display and make summary statements about data, through to highly detailed technical ones like implementation and assessment of causal models of the problems. This article is at the elementary end (the appropriate place to start) and introduces a) issues concerned with experimental design and choice of test data, b) definition of statistical terms involved with inference of the sample mean from the population mean. We start by outlining our motivation in initiating this series of articles and then emphasize how this article is different from a textbook on the subject.

Our experience in teaching and supervising projects with some trainee professionals who will deliver speech and hearing therapy to clients has shown that difficulties are experienced with statistics. It is our belief that part of the reason for this is that the texts and approaches often appear to be remote from the concerns of the students. In particular, texts for teaching students are often from other application areas about the study of behaviour (mainly psychology) that, though closely allied to speech and hearing science, are not directly applicable. When we have used examples from the speech and hearing sciences in our teaching, we have found that students seem able to access the material more readily. Even though we have limited ourselves to elementary statistics, the treatment is not comprehensive. There are two major omissions from these ‘elements’. The first concerns depiction of data. The second is statistical treatment when only a single case is available. It is hoped that both these omissions will be the subject of a future article.

The article is not a précis of a statistics course. There are many good such texts and other resources around (see the bibliography). These books mostly contain detailed information about how to describe data (only touched on here), hypothesis testing and cover the logic and mathematics behind different tests. The article should be used as a gateway to use one of the statistical techniques that give details of specific tests and procedures (though a couple of tests are considered to convey some important details). We do not mean to imply that all available textbooks lack such an overview. However, as textbooks are more extensive than this brief article, there is a tendency for this information to be distributed widely making it hard for students to relate it together. We envisage students reading this article and, once they have understood the material, being able to use it to access the appropriate information in other texts.

2. Experimental design and choice of data

When talking about procedural considerations in statistical analysis, it may help to make things concrete from the start by using an example drawn from hearing science. This is intended to introduce some concepts which are needed both for a researcher to evaluate studies and to allow an experimenter to do their own statistics. Let us assume that a company has developed a hearing aid (Aid A). It is to be employed in a country where all the inhabitants who have a hearing impairment might want to use it. The company wants some idea about how its performance compares with another aid on the market (Aid X). The company calls on a hearing clinic to evaluate the device and the clinic decides to assess it using read texts that can be carefully controlled, even though the aid will eventually have to operate with spontaneous speech to be useful for clients. Some of the questions the assessment team commissioned to do the work may decide to address are:

1. How to check whether there are differences between spontaneous and read speech, then make a decision whether the results with read speech apply to spontaneous speech.
2. If they find differences between read and spontaneous speech that require them to use the latter, how can they check whether language statistics on a sample of recordings are representative of the language as a whole?
3. How are appropriate test data chosen?

These points highlight some of the statistical analysis issues that feature in everyday clinical decisions. Similar questions would arise for a speech therapist when considering whether some new treatment they have learned about produces improvement in speech control so they can decide whether it is worth changing from the current procedure they deliver. In each case, the speech or hearing professional may or may not conduct the research themselves. When they rely on someone else's published report about the speech or hearing procedure, they need to know whether the analysis was conducted properly and how to interpret the results. Moreover, the specific questions raised, though pertaining to a particular issue of concern, are illustrative of many similar problems that clinicians encounter. Now we will set about attempting to provide answers to these (and other) questions.

Statistical and experimental procedures for analyzing data

In the first part of this section, some fundamental ideas in statistics will be illustrated through selected examples drawn from the speech and hearing sciences.

Statistics is the acquisition, handling, presentation and interpretation of numerical data. Speech and hearing scientists have considerable experience acquiring, handling, presenting and interpreting communication data.

Populations, samples and other terminology

A *population* is usually defined as the collection of units to which inference (from the sample) is desired (the units may be people, phonemes etc.). In the earlier example, all hearing impaired individuals in the country are the population. Here everyday use of the term 'population' corresponds with its use in a statistical sense. Though population in a statistical sense can have the same meaning as the geographical sense, it need not be the case. Thus, for instance the population of users of a hearing aid specifically developed for presbyacusis would only comprise individuals with this disorder. Population does not only refer to humans - for example, the population of /p/ phonemes of a speaker would be all of the instances of that phoneme a speaker ever produces.

A *variable* ranges over numerical values associated with each unit of the population. Variables are classed as either *independent* or *dependent variables*. An independent (or, as some statisticians prefer, explanatory) variable is one that is controlled or manipulated by the researcher. So, for example, when setting up a test for a hearing aid or some treatment for a speech problem, the experimenter might consider it necessary to ensure that as many females are recorded in the test data as males. Sex would then be an independent variable (independent variables are also referred to as factors, particularly in connection with the statistical technique *Analysis of Variance*, *ANOVA* discussed in a later section). A dependent variable is a variable that the investigator measures to determine the effect of the independent variable.

When a variable is measured on all units of a population, a full census has been taken. If it were always possible to obtain census data, there would be no need for inferential statistics. However, since most speech and hearing applications (and, indeed, in many other aspects that require measurement), involve very large or infinite populations (such as those illustrated earlier of speakers or phonemes), it is not possible to measure variables on all units: In these circumstances, a finite sample is taken. This sample is used to study the variable of concern in the population. So, if you wanted an idea of the average voice fundamental frequency of men, you might make measurements on a sample of 100 men. This sample is then studied as if it is representative of the population. The statistician is able to provide information about the relationship between variables measured on the sample (here its mean) and, what the investigator is really interested in, the mean voice fundamental frequency of the population.

Sampling

The main problem in treating data statistically is how to ensure the reliability of information about the population obtained from a sample. The main requirement to achieve this is to take a simple random sample. A sample is *simple random* if every member of the population has the same chance of being selected as every other member. Thus, if some voice recognition equipment is needed for a research application, and a check is made on its performance using employees from the clinic, the sample would not be simple random: It is unlikely that the employees from the clinic are from all social strata, there may be gender imbalances, and they would only include people of working-age.

Biases

Selection of a sample that is not a simple random sample is one of the main sources of bias in assessments. Bias can be defined as a systematic tendency to misrepresent the population. So, if the speech recognizer mentioned in the previous section is intended to be used by all members of the population, you cannot select an unbiased sample of speakers from a sample of people recorded just between 9 a.m. and 5 p.m. This would exclude people who are at work; so, if you do this, the result is a biased sample which is not necessarily representative of the target population.

If you take a sample, how sure can you be that if you measure variables such as the mean of the sample is close to the mean of the population? This sort of problem is termed *estimation* and is considered in the following.

Estimating sample means, proportions and variances

Estimation is used for making decisions about populations based on simple random samples. A truly random sample is likely to be representative of the population; this does not mean that a variable measured on a second sample taken will be the same as the first. The skill involved in estimating the value of a variable is to impose conditions which allow an acceptable degree of error in the estimate without being so conservative as to be useless in practice (an extreme case of the latter would be recommending a sample of the same order of magnitude as the population). The necessary background skill is to understand how quantities like sample means, proportions and variances are related to means, proportions and variances in the population. The following notation is used in the discussion: M is the sample mean, S is the sample standard deviation, and S^2 is the sample variance; μ, is the population mean, σ is the population standard deviation, and σ^2 is the population variance. The abbreviations *sd* and S.D. are sometimes used for standard deviation; S.E. is used for standard error, Z is used for z-scores, >> p_{est} << stands for estimated probability, and p stands for proportion.

Estimating means

A fundamental step towards this goal is to relate the sample statistic to a probability distribution: What this means is: if we repeatedly take samples from a population, how do the variables measured on the sample relate to those of the population? To translate this to an empirical example: How sure can you be about how close your sample mean lies to the population mean? Even more concretely, if we obtain the mean of a set of samples, how does the mean of a particular sample relate to the mean of the population? As has already been said, the value of the mean of the first of two samples is unlikely to be exactly the same as the second. However, if repeated samples are taken, the mean value of all the samples will cluster around the population mean; this is usually regarded as an unbiased estimator of the population mean.

The usual way this is shown in textbooks is to take a known distribution (i.e., where the population mean is known) and then consider what the distribution would be like when samples of a given size are taken. So, if a population of events has equally likely outcomes and the variable values are 1, 2, 3 and 4, the mean would be 2.5. If all possible combinations are taken (1 and 2, 1 and 3, 1 and 4, 2 and 3, 2 and 4, 3 and 4), the mean of the mean values for all pairs is also 2.5 (taking all pairs is a way of ensuring that the sample is simple random). An additional important finding is that if the distribution of sample means (the sampling distribution) are plotted as a histogram, the distribution is no longer rectangular (rectangular because each option was assumed to occur with the same frequency) but has a peak at 2.5 (1 and 4, and 2 and 3 both have a mean of 2.5 and none of the means of the other pairs have the same value, thus, the peak at 2.5). Moreover the distribution is symmetrical about the mean and approximates more to a normal (Gaussian) distribution even though the original distribution was not. As sample size gets larger the approximation to the normal distribution gets better. Moreover, this tendency applies to all distributions, not just the rectangular distribution considered. The tendency of large samples to approximate the normal distribution is, in fact, a case of the *Central Limit Theorem*.

This particular result has far-reaching implications when testing between alternative hypotheses (see below). As a rule of thumb sample sizes of 30 or greater are adequate to approximate the mean of a normal distribution (though this depends on the nature of the parent distribution).

The statistical quantity *standard deviation* (S) is a measure of how a set of observations $x_1 - x_n$ where n is the number of observations scatter about the mean (x with the bar above it). It is defined numerically as:

$$S = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Later the related quantity of the variance will be needed. This is simply the *sd* squared:

$$S^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

An important aspect of the situation described is that the sample means themselves (rather than the observations) have a standard deviation (*sd*). The *sd* of the sample means (here the *sd* of all samples of size two for the rectangular distribution) is related to the *sd* of the samples in the original distribution by the formula:

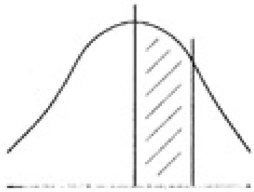
$$S.E. = \frac{\sigma x}{\sqrt{n}}$$

This quantity is given a particular name to distinguish it from the *sd* - it is called the *standard error* (*S.E.*). In practice, the standard deviation of the population is often not known. In these circumstances, provided the sample is sufficiently large, the standard deviation of the sample can be used to approximate that of the population and the above formula used to calculate the S.E. The S.E. is used in the computation of another quantity, the z score of the sample mean:

$$Z = \frac{\bar{x} - \mu x}{S.E.}$$

The importance of this quantity is that the measure can be translated into a probabilistic statement relating the sample and population means. Put another way, from the *z* score, the probability of a sample mean being so far from the population mean can be computed.

To show how this is used in practice: if a sample of size 200 is taken, what is the probability that the mean is within 1.5 S.E.s of the population mean? Normal distribution tables give the desired area. Here is a section of a table giving the proportion of the area of a normal distribution associated with given values of *z* (the stippled section in the figure indicates what area is tabulated) where the mean is the line at the peak and the line to its right is 1.5 S.E. away:



z	Area
...	...
1.3	0.4032
1.4	0.4192
1.5	0.4332
1.6	0.4452
...	...

The sketch of the normal distribution is symmetrical and the symmetry is about the mean value (i.e., the peak of the distribution). The z values above the mean are tabulated, and the row with a z value of 1.5 indicates that 0.4332 of the area on the right half of the distribution lie within 1.5 S.E.s above the mean. Since it has already been noted that the distribution is symmetrical, 0.4332 of the area will lie within 1.5 S.E.s below the mean. Thus, the area within 1.5 S.E.s above or below the mean is 0.4332 + 0.4332, or 0.8664. Thus, converted to percentages, approximately 86.6 of all samples of size 200 will have means within 1.5 S.E.s of the population mean. If, as in any real experiment, one sample is taken, we can assign a statement about how likely that sample is being within the specified distance of the mean.

Another, related, use of S.E.s is in stipulating confidence intervals. If you look at the areas associated with particular z values in the way just described, you should be able to ascertain that the area of a normal distribution enclosed within *z* values ±1.96 S.E.s of the mean M is 95. Thus, if the S.E. and M of a sample are known, you can specify a measurement interval that indicates the degree of confidence (here 95%) that the population mean will be within these bounds. This is between the value 1.96 times the S.E below the sample mean and 1.96 times the S.E. above the sample mean. This particular interval is called the 95% confidence interval. Other levels of confidence can be adopted by obtaining the corresponding *z* values.

Since this topic is so important, an example is given: Say a random sample of mean voice fundamental of 64 male university students has a mean of 98 Hz and a standard deviation of 32 Hz. What is the 95% confidence interval for mean voice fundamental of the male students at this university? The maximum error of the estimate is approximated (using sample standard deviation S rather than that of the population σ as an approximation, see above) as:

$$1.96 \cdot \frac{S}{\sqrt{n}} = 1.96(\frac{32}{\sqrt{64}}) = 7.84$$

Thus, the 95% confidence interval is from 98 - 7.84 = 90.16 to 98 4- 7.84 = 105.84. Often, the confidence intervals are presented graphically along with the means: the mean of the dependent variable is indicated on the *y* axis with some chosen symbol; a line representing the confidence interval extends from (in this case) 90.16 to 105.84 and it is drawn vertically and passes through the mean.

Before leaving this section, it is necessary to consider what to do when wanting to make corresponding statements about small-sized samples which cannot be approximated with the normal distribution. Here computation of the mean and standard error proceeds as before. Since the quantity *z* is used in conjunction with the normal distribution tables, it cannot be used. Instead the analogous quantity *t* is calculated:

$$t = \frac{\bar{x} - \mu x}{S.E.}$$

The distribution of *t* is dependent on sample size *n* and so (in essence) the *t* value has to be referred to different tables for each size of sample. The tables corresponding to the *t* distribution are usually collapsed into one table and the section of the table used is accessed by a parameter related to the sample size *n* (the quantity used for accessing the table is *n - 1* and is called the degrees of freedom). Clearly, since several different distributions are being tabulated, some condensation of the information relative to the *z* tables is desirable. For this reason, *t* values corresponding to particular probabilities are given. Consideration of *t* tables emphasizes one of the advantages of the Central Limit Theorem insofar as one table can be used to address a wide variety of issues rather than is the case for *t*.

Estimating proportions

Here the problem faced is similar to that with means: A sample has been taken and the *proportion* of people meeting some criterion and those not meeting that criterion are observed. The question is with what degree of *confidence* can you assert that the proportions observed reflect those in the population? Once again the solution is directly related to that discussed when estimating how close a sample mean lies to the population mean using z scores. Essentially the *z* score for means measures:

$$Z = \frac{\textit{estimated mean} - \textit{population mean}}{S.E.}$$

The only difference here is that binomial events are being considered (meet/not meet the criterion). The z score associated with a particular sample based on the estimated probability and the population proportion is (where q = 1-p): >>phat<<

$$Z = \frac{\hat{P} - p}{\sqrt{(pq/n)}}$$

Normal distribution tables can again be used to assign a probability associated with this particular outcome.

To illustrate with an example: Suppose that it is expected that as many men will use the speech recognizer as will women (p (man) = p (woman) 0.5). What size of sample is needed to be 95% certain that the proportion of men and women in the sample differs from that in the population by at most 4%?

$$1.96(= 95\%) = \frac{.04}{\sqrt{(.5)(.5)/n}}$$

Solving for *n* gives 600.25. Therefore, a sample of size at least 601 should be used.
Now what are the effects if we want to be more than 4 confident, say if the difference is reduced to 2%. The required sample size jumps to 2401.

Estimating variance

The relationship between the variance of a sample and that of the population is distributed as χ^2 (chi squared) with *n - 1* degrees of freedom.

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

Thus, if we have a sample of size 10 drawn from a normal population with population variance 12, the probability of its variance exceeding 18 is:

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

$$= \frac{9 \cdot 18}{12} = 13.5$$

This has associated with it 9 degrees of freedom. Because χ^2 values are only tabulated for particular probabilities (as with *t*), the probability can only be estimated for limited probabilities. In this case $\chi 2$ lies between 0.2 and 0.1.

Ratio of sample variances

If two independent samples are taken from two normal populations with variance σ_1^2 and σ_2^2, the ratio of the two variances (S_1^2 and S_2^2) has the *F* distribution:

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

If the two samples (which can differ in size) from the same normal population are taken, then the ratio of the variances will be approximately 1. Conversely, if the samples are not from the same normal population, the ratio of their variances will not be 1 (the ratio of the variances is termed the *F ratio*). The *F* tables can be used to assign probabilities that the sample variances were or were not from the same normal distribution. The importance of this in the Analysis of Variance (ANOVA) will be seen later.

3. Statistical terms involved in inference to the population mean from the sample mean

Simple hypothesis testing

Many practical and research applications in speech and hearing science require testing of hypotheses. An example from the scenario given at the outset was testing whether there were differences between read and spontaneous speech with respect to selected statistics. If the statistic was mean vowel duration in the two conditions where speech was recorded, we have a situation calling for simple hypothesis testing. This situation is called simple hypothesis testing since it involves a parameter of a single population.

Following the approach adopted so far, the concepts involved in such testing are illustrated for this selected example. The first step is to make alternative assertions about what the likely outcome of an analysis might be. One assertion is that the analysis might provide no evidence of a difference between the two conditions. This case is referred to as the null hypothesis (conventionally abbreviated as Ho) and might assert here that the mean syllable duration in the read speech is the same as that in the spontaneous speech. Other assertions might be made about this situation. These are referred to as *alternative hypotheses*. One alternative hypothesis would be that the vowel duration in the read speech will be less than that of the spontaneous speech. A second would be the converse, i.e. the vowel duration in the spontaneous speech will be less than that of the read speech. The decision about which of these alternate hypotheses to propose will depend on factors that lead the speech or hearing student or investigator to expect differences in one direction or the other. These instances are referred to as *one-tailed (one-directional) hypotheses* as each predicts a specific way in which there will be a difference between read and spontaneous speech. If the investigator wants to test for a difference but has no theoretical or empirical reasons for predicting the direction of the difference, then the hypothesis is said to be two-tailed. Here, large differences between the means of the read and spontaneous speech, no matter which direction they go in, might constitute evidence in favor of the alternative hypothesis.

The distinction between one and two-tailed tests is an important one as it affects what difference between means is needed to assert a significant difference (i.e., support the null hypothesis). In the case of a one-tailed test, smaller differences between means are needed than in the case of two-tailed tests. Basically, this comes down to how the tables are used in the final step of assessing significance (see below). There are no fixed conventions for the format of tables for the different tests, so there is no point in illustrating how to use them. The tables usually contain guidance as to how they should be used to assess significance.

Hypothesis testing involves asserting what level of support can be given in favor of, on the one hand, the null, and, on the other, the alternate hypotheses. Clearly no difference between the means of the read and spontaneous speech would indicate that the null hypothesis is supported for this sample. A big difference between the means would seem to indicate that there is a statistical difference between these samples if the direction in which the means differs is in the same direction as hypothesized for a one-tailed hypothesis or if a two-tailed test has been formulated. The way in which a decision whether a particular level of support (a probability) is provided is described next.

In the read-spontaneous example that we have been working through, we are interested in testing for a difference between means for two samples where, it is assumed, the samples are from the same speaker. The latter point requires that a related groups test as opposed to an independent groups test is used. In this case, the *t* statistic is computed from:

$$t = \frac{\text{mean of condition}_1 - \text{mean of condition}_2}{S.E.\text{ of differences}}$$

Thus if the read speech for 15 speakers had a mean vowel duration of 40.2 milliseconds and the spontaneous speech 36.4 milliseconds and the standard deviation of the difference between the means is 2.67, the *t* value is 1.42. The *t* value is then used for establishing whether two sample means lying this far apart might have come from the same (null hypothesis) or different (alternate hypothesis) distributions. This is done by consulting tables of the *t* statistic using *n-l* degrees of freedom (here *n* refers to the number of pairs of observations).

In assessing a level of support for the alternate hypothesis, decision rules are formulated. Basically this involves stipulating that if the probability of the means lying this far apart is so low then a more likely alternative is that the samples are drawn from different populations, assuming that the samples are from the same distribution. The "stipulation" is done in terms of discrete probability levels and, conventionally, if there is a less than 5% chance that the samples were from the same distribution, then the hypothesis that the samples were drawn from different distributions is supported (the alternative hypothesis at that level of significance). Conversely, if there is a greater than 5 in a hundred chance that the samples are from the same distribution, the null hypothesis is supported. In the worked example, with 14 degrees of freedom, a *t* value of 1.42 does not support the hypothesis that the samples are drawn from different populations, thus the null hypothesis is accepted. It should be noted that support or rejection of these alternative hypotheses is statistical rather than absolute. In 1/20 (5%) cases where no difference is asserted, a difference does occur (referred to as a Type II error, accepting the null hypothesis when it is false) and in cases where a 5% significance level is adopted and differences found, 1 occasion out of 20 will also lead to an error (referred to as a Type I error, rejecting the null hypothesis when it is in fact true).

Analysis of Variance

As was said earlier, this is not supposed to be a substitute for your statistics textbook as, in particular, it does not cover all statistical tests that might be encountered. It only offers an overview and a means of accessing relevant material in a textbook.

However, some comments on Analysis of Variance (ANOVA) are called for as it is a technique that has a widespread use in speech and hearing assessment. ANOVA is a statistical method for assessing the importance of factors that produce variability in responses or observations. The approach is to control for a factor by specifying different values (or, treatment levels) for it in order to see if there is an effect. It can be thought of as having sampled a potentially different population (different in the sense of having different means). Factors that have an effect change the variation in sample means, where "factor" refers to a controlled independent variable. When the experimenter controls the levels of the factors, this is referred to as *treatment level*.

For example, in the ANOVA approach, two estimates of the variances are obtained: the variance between the sample means, between groups variance, and the variance of each of the scores about their group mean, within groups variance. If the treatment factor has had no effect, then variability between and within groups should both be estimates of the population variance. So, as discussed earlier when the ratio of two sample variances from the same population was considered, if the ratio of between groups to within groups is taken, the value should be about 1 (in which case, the null hypothesis is supported). The ratio of two variances is called the *F* ratio. Statistical tables of the *F* distribution can be consulted to ascertain whether the *F* ratio is large enough to support the hypothesis that the treatment factor has had an effect resulting in larger variance of the between group to the within group means (the alternative hypothesis is supported). Another way of looking at this is that the between groups variance is affected by individual variation of the units tested plus the treatment effect whereas the within groups estimate is only affected by individual variation of the units tested.

ANOVA is a powerful tool which has been developed to examine treatment effects involving several factors. Some examples of its scope are that it can be used with two or more factors. Factors that are associated with independent and related groups can be tested in the same analysis, and so on. When more than one factor is involved in an analysis, the dependence between factors (interactions) comes into play and has major implications for the interpretation of results.

Non-parametric tests

Parametric tests cannot be used when discrete, rather than continuous, measures are obtained since the Central Limit Theorem does not approximate the normal distribution in these instances. The distinction between discrete and continuous measures is the principal factor governing whether a parametric or non-parametric test can be employed. Continuous and discrete measures relate to another taxonomy of scales - interval, nominal and ordinal: interval scales are continuous and the others are discrete. Statisticians consider this taxonomy misleading, but since it is frequently encountered in the behavioral sciences in general, the nature of data from the different scales is described. Interval data are obtained when the distance between any two numbers on the scale are of known size and is characterized by a constant unit of measurement. This applies to physical measures like duration and frequency measured in Hertz (Hz) which have featured in the examples discussed to now. Nominal scales are obtained when the measures are obtained from symbols to characterize objects (such as sex of the speakers). Ordinal scales give some idea of the relative magnitude of units that are measured but the difference between two numbers does not give any idea of the relative size. Examples would be responses to questionnaires where there is no guarantee of equal distance between the response choices offered (e.g. strongly agree, agree, disagree, strongly disagree). In cases where parametric tests cannot be used, non-parametric (also known as distribution-free) tests have to be employed. The computations involved in these tests are straightforward and covered in any elementary text book. A reader who has followed the material presented thus far should find it easy to apply the previous ideas to these tests.

A number of representative questions a speech and hearing investigator might want to answer were considered at the start of this section. Let us just go back over these and consider which ones we are now equipped to answer. First there was how to check whether there are differences between spontaneous and read speech.

If the measures are parametric (such as would be the case for many acoustic variables), then either an independent or related *t* test would be appropriate to test for differences. An independent *t* test is needed when samples of spontaneous speech and read speech are drawn from different speaker sets; a related *t* test is used when the spontaneous and read samples are both obtained from the same group of speakers.

If the measures are non-parametric (e.g. ratings of clarity for the spontaneous and read speech) then a *Wilcoxon test* would be used when the read and spontaneous versions of the speech are drawn from the same speaker and a *Mann-Whitney U test* otherwise.

If you find differences between read and spontaneous speech (see application described), how can you check whether language statistics on your sample of recordings is representative of the language as a whole - or, what might or might not be the same thing, how can you be sure that you have sampled sufficient speech? For this, the background information provided to estimate how close sample estimates are to population estimates is appropriate.

4. Conclusions

This has been a whirlwind tour of some elementary concepts in treatment of data from the speech and hearing areas starting with issues concerned with choice of data and inferential statistics. A final aim is to draw readers attention to other simple, clearly written resources for examining data in these (and medical) areas in general. For this purpose, a selected bibliography follows.

Acknowledgement. The second author is supported by the Wellcome Trust grant 072639.

Bibliography

Armitage, P., Berry, G., & Matthews, J.N.S. (2001). *Statistical Methods in Medical Research*. London: Blackwell Publishers. *A standard text on statistical methods in Medical Research.*

Bland, M. (2000). *An introduction to Medical Statistics, 3rd edition*. Oxford: Oxford University Press.. Another standard text on statistical methods in Medical Research.

Hand, D. J. (2004). Measurement - theory and practice. London: Hodder Arnold. *Just out covers elementary and advanced levels.*

Huff, D (1993). How to lie with statistics. New York: Norton publishers. *Popular and accurate*.

Moser, C & Kalton, G. (1979), *Survey methods in social investigation*. Heinemann Educational: London

British Medical Journal: Statistics Notes

Perhaps the finest series of short articles on the use of statistics is the occasional series of Statistics Notes started in 1994 by the British Medical Journal. It should be required reading in any introductory statistics course. The full text of the articles is available is available on the World Wide Web. The articles are listed here chronologically.

- Correlation, regression, and repeated data J Martin Bland & Douglas G Altman BMJ 1994;308:896 (2 April)
- Regression towards the mean J Martin Bland & Douglas G Altman BMJ 1994;308:1499 (4 June)
- Diagnostic tests 1: sensitivity and specificity Douglas G Altman & J Martin Bland BMJ 1994;308:1552 (11 June)
- Diagnostic tests 2: predictive values Douglas G Altman & J Martin Bland BMJ 1994;309:102 (9 July)
- Diagnostic tests 3: receiver operating characteristic plots Douglas G Altman & J Martin Bland BMJ 1994;309:188 (16 July)
- One and two sided tests of significance J Martin Bland & Douglas G Altman BMJ 1994;309:248 (23 July)
- Some examples of regression towards the mean J Martin Bland & Douglas G Altman BMJ 1994;309:780 (24 September)
- Quartiles, quintiles, centiles, and other quantiles Douglas G Altman & J Martin Bland BMJ 1994;309:996 (15 October)
- Matching J Martin Bland & Douglas G Altman BMJ 1994;309:1128 (29 October)
- Multiple significance tests: the Bonferroni method J Martin Bland & Douglas G Altman BMJ 1995;310:170 (21 January)
- The normal distribution Douglas G Altman & J Martin Bland BMJ 1995;310:298 (4 February)
- Calculating correlation coefficients with repeated observations: Part 1--correlation within subjects J Martin Bland & Douglas G Altman BMJ 1995;310:446 (18 February)
- Calculating correlation coefficients with repeated observations: Part 2--correlation between subjects J Martin Bland & Douglas G Altman BMJ 1995;310:633 (11 March)
- Absence of evidence is not evidence of absence Douglas G Altman & J Martin Bland BMJ 1995;311:485 (19 August)
- Presentation of numerical data J Martin Bland & Douglas G Altman BMJ 1996;312:572 (2 March)
- Logarithms J Martin Bland & Douglas G Altman BMJ 1996;312:700 (16 March)
- Transforming data J Martin Bland & Douglas G Altman BMJ 1996;312:770 (23 March)
- Transformations, means, and confidence intervals J Martin Bland & Douglas G Altman BMJ 1996;312:1079 (27 April)
- The use of transformation when comparing two means J Martin Bland & Douglas G Altman BMJ 1996;312:1153 (4 May)
- Comparing several groups using analysis of variance Douglas G Altman & J Martin Bland BMJ 1996;312:1472-1473 (8 June)

- Measurement error and correlation coefficients J Martin Bland & Douglas G Altman BMJ 1996;313:41-42 (6 July)
- Measurement error proportional to the mean J Martin Bland & Douglas G Altman BMJ 1996;313:106 (13 July)
- Interaction 1: heterogeneity of effects Douglas G Altman & John NS Matthews BMJ 1996;313:486 (24 August)
- Measurement error J Martin Bland & Douglas G Altman BMJ 1996;313:744 (21 September)
- Interaction 2: compare effect sizes not P values John NS Matthews & Douglas G Altman BMJ 1996;313:808 (28 September)
- Interaction 3: How to examine heterogeneity John NS Matthews & Douglas G Altman BMJ 1996;313:862 (5 October)
- Detecting skewness from summary information Douglas G Altman & J Martin Bland BMJ 1996;313:1200 (9 November)
- Cronbach's alpha J Martin Bland & Douglas G Altman BMJ 1997;314:572 (22 February)
- Units of analysis Douglas G Altman & J Martin Bland BMJ 1997;314:1874 (28 June)
- Weighted comparison of means J Martin Bland & Sally M Kerry BMJ 1998;316:129 (10 January)
- Sample size in cluster randomisation Sally M Kerry & J Martin Bland BMJ 1998;316:549 (14 February)
- The intracluster correlation coefficient in cluster randomisation Sally M Kerry & J Martin Bland BMJ 1998;316:1455-1460 (9 May)
- Generalisation and extrapolation Douglas G Altman & J Martin Bland BMJ 1998;317:409-410 (8 August)
- Time to event (survival) data Douglas G Altman & J Martin Bland BMJ 1998;317:468-469 (15 August)
- Bayesians and frequentists J Martin Bland & Douglas G Altman BMJ 1998;317:1151-1160 (24 October)
- Survival probabilities (the Kaplan-Meier method) J Martin Bland, & Douglas G Altman BMJ 1998;317:1572-1580 (5 December)
- Treatment allocation in controlled trials: why randomise? Douglas G Altman & J Martin Bland BMJ 1999;318:1209-1209 (1 May)
- Variables and parameters Douglas G Altman & J Martin Bland BMJ 1999;318:1667-1667 (19 June)
- How to randomise Douglas G Altman & J Martin Bland BMJ 1999;319:703-704 (11 September)
- The odds ratio Douglas G Altman & J Martin Bland BMJ 2000;320:1468 (27 May)
- Blinding in clinical trials and other studies Simon J Day & Douglas G Altman BMJ 2000;321:504 (19 August)
- Concealing treatment allocation in randomised trials Douglas G Altman & Kenneth F Schulz BMJ 2001;323:446-447 (25 August)
- Validating scales and indexes J Martin Bland & Douglas G Altman BMJ 2002;324:606-607 (9 March)
- Interaction revisited: the difference between two estimates Douglas G Altman & J Martin Bland BMJ 2003;326:219 (25 January)
- The logrank test J Martin Bland & Douglas G Altman BMJ 2004 328(7447):1073 (1 May)

TARGET ARTICLE

The Use of Structural Equation Modeling in Stuttering Research: Concepts and Directions

Stephen Z. Levine
Adjunct Lecturer, Department of Behavioral Sciences
Beer-Sheva University, Israel
szlevine@technion.ac.il

K. V. Petrides
School of Psychology and Human Development
Institute of Education, University of London, UK
k.petrides@ioe.ac.uk

Stephen Davis
Department of Psychology, Centre for Human Communications, Institute of Cognitive Neuroscience, and Institute of Movement Neuroscience, University College London, Gower St., London WC1E 6BT, England
stephen.davis@ucl.ac.uk

Chris J. Jackson
School of Psychology
University of Queensland, Brisbane, Australia
c.jackson@psy.uq.edu.au

Peter Howell
Department of Psychology, Centre for Human Communications, Institute of Cognitive Neuroscience, and Institute of Movement Neuroscience, University College London, Gower St., London WC1E 6BT, England
p.howell@ucl.ac.uk

Abstract. This article provides a brief introduction to the history and applications of the class of data analytic techniques collectively known as Structural Equation Modeling (SEM). Using an example based on psychological factors thought to affect the likelihood of stuttering, we discuss the issues of specification, identification, and model fit and modification in SEM. We also address points relating to model specification strategies, item parceling, advanced modeling, and suggestions for reporting SEM analyses. It is noted that SEM techniques can contribute to the elucidation of the developmental pathways that lead to stuttering. **Keywords:** Structural equation modeling, LISREL, stuttering models.

1. Introduction I. Multifactor models and stuttering

Without going into the nuances of the many definitions of stuttering that are available, the disorder involves difficulty in handling language and is usually manifest as problems in overt speech control. This does not necessarily mean that language problems are paramount and the sole cause of stuttering. The anxiety a speaker is experiencing, the situation in which he or she is interacting and so on will determine whether or not the speaker stutters at that particular time.

The aim of this paper is to establish how the multiple and various factors associated with stuttering behavior, work together in influencing stuttering. Before introducing the Structural Equation Modeling (SEM) technique, it is necessary to look at some of the issues and debates that have been raised in connection with the process of theory construction in the field of stuttering. This shows that more detailed specification of theories is needed before SEM can be employed as a way of implementing, assessing and evaluating alternative multifactor models of stuttering (the intention is that this article will stimulate work on both theory construction and evaluation).

The first thing to note is that there are several models of stuttering that are specified in detail that have been tested experimentally. In the main, these have been developed to try to explain how language influences speech-motor performance and how this then results in stuttering. Kolk and Postma's (1997) covert repair hypothesis (CRH) offers one example. This explains stuttering in terms of a linguistic system that can malfunction and produce errors which then affect speech-motor performance. Stuttering, according to CRH, can be characterized as the response an individual makes to linguistic errors, which produces different types of stuttering events (such as repetitions and prolongations) in overt speech (Smith, 1999). Put another way, stuttering is viewed by CRH as a linear causation process in which errors in the linguistic processes lead directly to observed problems in speech output (Smith, 1999). CRH does not concern itself with other factors that can influence stuttering. It is simply an account of the particular link between language processes and speech output behaviour. In tests that have been made of CRH, other factors that affect stuttering would have minimal influence provided good experimental practice is followed. An example of such practices can be seen for gender, which is known to affect stuttering behaviour, and language more generally. Any influences of gender per se can be controlled by selecting speakers of the same sex for both stuttering and control groups. It is not that CRH maintains that gender or other factors are not relevant to stuttering, simply that with suitable precautions, authors approaching their research from CRH's perspective are able to focus on what interests them – the link between language processes and speech performance.

Second, also as a consequence of using experimental control principles, the problem of the relationship between language and speech output is simplified into one of linear causation. This does not imply that if other factors are added to the model, they too have to be placed at some point in a linear chain. If, for instance, CRH was extended to account for the effects of anxiety on stuttering, anxiety might influence the linguistic system directly by, for instance, increasing the likelihood that a speaker who stutters makes more errors in that system. Alternatively, anxiety might be one factor that increases overall arousal level, making speakers attempt utterances faster than they otherwise might. This could then, in turn, put pressure on both the linguistic and motor systems (if the linguistic system operates more rapidly, it may make more errors, Kolk and Postma, 1997; if the motor system executes speech plans more rapidly, it may show more variable performance, Smith and Kleinow, 2000). The arousal system might then be modeled as an independent parallel process that has indirect influences on both language and speech processes. If anxiety level is controlled, this nonlinear causal chain is reduced to a linear one for the purpose of study.

The third issue is whether there is only one appropriate measure of output behavior (the dependent variable). CRH focuses on events that are regarded as being associated with stuttering, such as the repetitions and prolongations mentioned earlier. These are suited to their authors' aims where they wish to examine specifically the influence of the linguistic system on failures in speech output. Smith uses an output that reflects the (possibly nonlinear) influences of a variety of additional factors. Others who are interested in the effects of a clinical procedure on speech outcomes might want to use a standardized instrument as their output measure (such as SSI-3, Riley, 1994).

Each of these output measures is appropriate for the particular research needs and no one measure meets all the requirements. This is not a point of view that is shared by all workers. Smith (1999), for instance, sees serious limitations in using speech event measures. One argument she puts forward is that studying stuttering by examining the overt manifestations of the problems (the stuttering events like part-word repetitions or prolongations) is analogous to studying volcanoes by examining the shape and type of effused eruptive material. In both these examples, the surface form (of language, and of the geological forms, respectively) is the output object. Smith points out that examination of surface forms of volcanoes did not advance ability to forecast eruptions. Plate tectonics provided the key to the underlying forces that allowed eruptions to be forecasted. The underlying dynamic movements of the plates lead to volcanoes at the points where they join. The analogy suggests that an underlying dynamic representation of speech (rather than events) might provide a unifying framework for understanding stuttering. This is not an argument we would buy (instead we prefer to retain our view that different output measures are suitable for different purposes). Moreover, we think the same may apply to the study of volcanoes: An important (and tried, tested and confirmed) prediction of plate tectonics is that volcanoes only occur at points where two plates abut. Clearly, some output measure that looked at the geographical distribution of volcanoes was needed to test this prediction.

This article is not concerned with issues in multifactor theory construction *per se*. It does seek to establish, once a multifactor theory has been specified, how a researcher can establish how well the theory (or model) fits the data and how that model can be compared with others. Modeling necessarily relies on having a theory specified in detail. Current efforts at theory construction in the stuttering area do not meet this requirement. For instance, Smith (1999, p.34) is a multifactor models that predict ways in which various component factors interact, but offer little in the way of specification of the exact form of that interaction (whether by direct or indirect influences as discussed in connection with anxiety). One major goal we hope to achieve is to stimulate authors to offer their theories in more detail (which would be welcomed as submissions to *Stammering Research*). The second major goal is to introduce SEM as an approach to discriminate between alternative theories. As there are no detailed and explicit multifactor theories at present, we use specified hypothetical models as the basis of our discussion. Before model specification, some general remarks about SEM are made.

2. Introduction II: Structural Equation Modeling and stuttering

Recent years have witnessed an unprecedented rate of growth in advanced quantitative techniques that are potentially applicable across a wide range of disciplines. With respect to non-experimental research in the social sciences, the development of new methods for data analysis has largely revolved around Structural Equation Modeling (SEM). Indeed, although SEM is mostly used for analysis of data collected by non-experimental methods, it is possible to conduct analyses of data obtained using experimental designs. Actually, SEM is a broad class of techniques covering confirmatory factor analysis, time growth analysis, multi-level latent modeling, and simultaneous equation modeling. The early origins of SEM can be traced back to the technique of path analysis, introduced in the field of biology by Wright (1934). However, the field of SEM as we now know it was created by the seminal contributions of Karl J. Jöreskog and his colleagues (notably Dag Sörbom), who together managed to integrate the data analytic traditions of multiple regression and factor analysis (e.g., Jöreskog, 1969, 1971). The breakthroughs in statistical theory were implemented in the computer program LISREL (LInear Structural RELations; e.g., Jöreskog & Sörbom, 1996), which has had a major impact across the social sciences and beyond.

The SEM literature has been growing rapidly over the last decade as reflected in the ever-increasing number of introductory chapters (e.g., Fife-Schaw, 2000), textbooks (e.g. Kline, 2004; Maruyama, 1998), and journal articles (e.g., Muthén, 2002). The results of a recent review of the relevant literature (Hershberger, 2003) showed that: (a) SEM has acquired dominance among multivariate techniques; (b) the number of journals publishing SEM articles continues to grow; and (c) '*Structural Equation Modeling: An Interdisciplinary Journal*' has become the primary outlet for the publication of technical developments in SEM. The journal contains tutorials by leaders in the field (e.g., Raykov & Marcoulides, 1999), and is a particularly useful resource for applied researchers, as is the online discussion group SEMNET (http://www.gsu.edu/~mkteer/semnet.html).

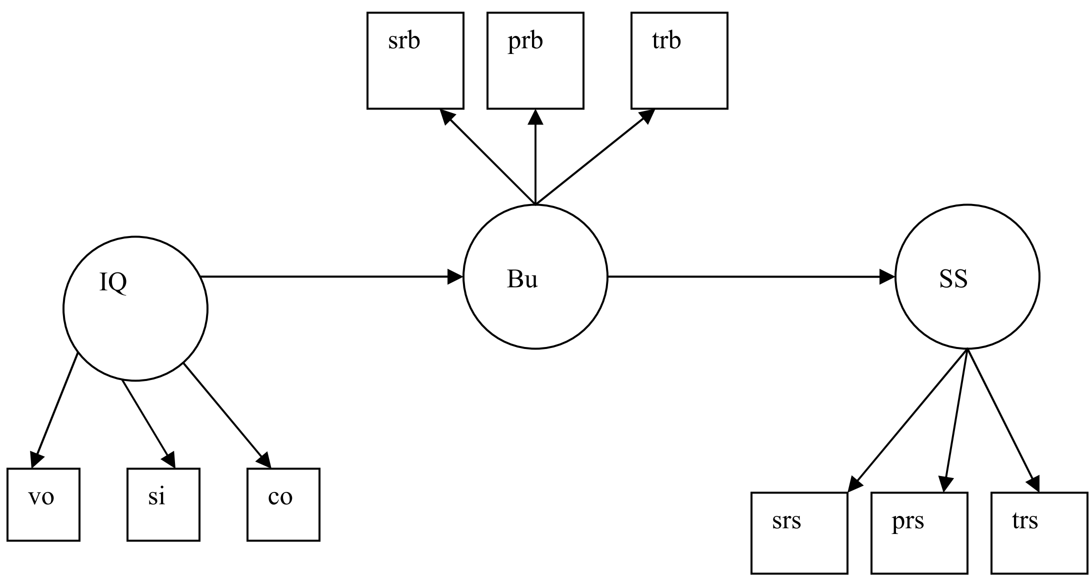
There are five main reasons for the recent surge in the popularity of SEM. First, it allows for statistical analyses that account for measurement error in the dependent and independent variables through the use of multiple observed indicators per latent variable. Second, it provides a rigorous approach to model testing, which requires careful theoretical development as discussed in the previous section. Third, it is extremely flexible as it can accommodate experimental designs, group differences designs, longitudinal designs, and multi-level designs, all of which are used in stuttering research. Fourth, it can accommodate tests of mediation and moderation (Baron & Kenny, 1986). Fifth, it allows for the statistical testing of multiple competing theoretical models (such as the two models proposed earlier about how anxiety could be added to CRH). While the typical approach to data analysis involves testing hypotheses based on a particular theory, a more powerful approach involves comparing alternative theories and establishing, based on *a priori* criteria and statistical indices, which model best accounts for the data. This elegant approach to developing theories and testing hypotheses has been applied in many different domains, including occupational psychology (e.g., Earley & Lituchy, 1991), personality (Petrides, Furnham, Jackson & Levine, 2003), criminological psychology (Levine & Jackson, 2004), and counselling psychology (Quintana & Maxell, 1999).

The main objective of this article is to introduce the new possibilities that SE modeling offers for the analysis of stuttering data. For the purposes of this exposition, we will employ a hypothetical example, illustrated in Figure 1 and based on socio-psychological variables that are implicated in stuttering (see Furnham & Davies, 2004). First some elementary steps that construct two theories of this process: Model 1 suggests that cognitive ability is influential in addressing whether a child is bullied which, in turn, influences the severity of stuttering (like the linear causation discussed in connection with anxiety and CRH). Model 2 differs in suggesting that cognitive ability directly influences the severity of stuttering, in addition to its indirect effect mediated via being bullied (in a similar way to the second alternative way of adding anxiety to CRH). In everyday terms, the first model maintains that a child with low IQ might be more prone to bullying, which then causes anxiety and leads to stuttering (linear causation model). The second model assumes IQ affects severity. IQ can also affect the degree to which a child is bullied that will then affect stuttering severity. SEM provides techniques that allow the best alternative model of the process involving these factors to be determined.

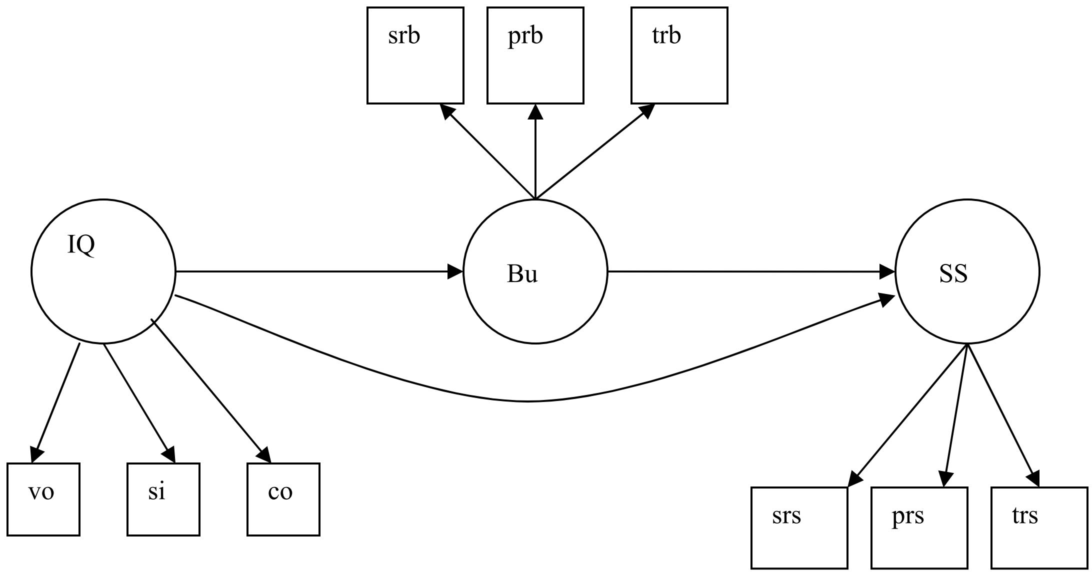
Path diagrams, like the one in Figure 1, are pictorial representations of an underlying system of mathematical equations. Observed indicators (items or scales) are designated by a box. In Figure 1, each latent variable (where a latent variable reflects the relationship between variables) is represented by three observed indicators in square boxes. Directional relationships are represented by straight lines with an arrowhead pointing toward the endogenous variable. Non–directional relationships (i.e., correlations) are represented by curved lines with arrowheads at both ends. Latent variables are represented by circles or ellipses in SEM and can be either exogenous or endogenous. *Exogenous* latent variables are 'upstream' variables that are not influenced by other variables in the model. *Endogenous* latent variables are 'downstream' variables that are influenced by other variables (exogenous, endogenous, or both) in the model. For instance, in Figure 1, IQ is an exogenous variable, whereas bullying and severity of stuttering are endogenous variables.

Figure 1. Two possible Competing Structural Models of Stuttering that propose a relationship between IQ, bullying and stuttering severity.

Model 1



Model 2



Note. IQ = General intelligence factor, vo = vocabulary test, si = similarities test, co = comprehension (example scales taken from WAIS – III tests). Bu = bullying, srb = self-reports bullying, prb = peer-reports of bullying, trb = teacher-reports of bullying. SS = Severity of stuttering, Srs = Self-reported severity of stuttering, prs = peer-reported severity of stuttering, trs = teacher-reports severity of stuttering.

Specification

The first stage in SEM involves specifying a model. A model is a statistical statement about the relationships between variables. Models take different forms, depending on the analytical approach that is adopted. In SEM, relationships are specified on the basis of the hypotheses outlined by theory. Specification refers to the translation of a theory into a structural model stating specific relationships between the variables. These relationships entail parameters that have magnitude and direction (+ / -) and can be either fixed or free. *Fixed parameters* are pre-specified by the researcher, rather than estimated from the data, and are usually, but not necessarily, set at zero. *Free parameters* are estimated from the data and are usually, again albeit not necessarily, significantly different from zero. In the example shown in Figure 1, the direct relationship between IQ and severity of stuttering is a fixed parameter, as it is set to zero in model 1 (i.e., no relationship is specified). In model 2, the relationship between IQ and severity of stuttering is a free parameter (i.e., estimated from the data).

The various parameters in an SEM define its two components, namely the measurement and the latent parts of the model. 'The *measurement model* is that component of the general model in which latent variables are prescribed. Latent variables are unobserved variables implied by the covariances among . . . [three] or more indicators' (Hoyle, 1995, p. 3). Latent variables are often termed factors and are free of the random error and uniqueness associated with their observed indicators. Scores on the observed indicators comprise common factor variance (i.e., variance shared with the other indicators of the construct), specific factor variance (i.e., reliable variance specific to the observed indicator), and measurement error. The specific variance and measurement error parts are collectively described as a variable's uniqueness.

In Figure 1 there are three observed indicators (represented by boxes) defining each of the three latent variables (represented by circles). It is noted that there are three observed indicators for each latent variable in Figure 1. The exact ratio of observed to latent variables remains a topic of debate, and appears to be somewhat dependent on the data at hand. In Figure 1, for example, the latent variable 'severity of stuttering' influences the observed indicators of self, peer, and teacher ratings of stuttering. The underlying rationale is that each of these ratings partially reflects the target's standing on the general latent variable of severity of stuttering.

One of the main strengths of the SEM approach is that it allows the decomposition of the relationships between variables. There are five different types of effects. *Direct effects* refer to direct relationships between variables, whereas *indirect effects* refer to relationships that are mediated via intervening variables.

In both models in Figure 1, IQ has a direct effect on bullying, which is a mediating variable that transmits the effect of IQ on severity of stuttering. In addition, it is possible for the relationship between two variables to be caused by a third variable, which is referred to as a *non-causal relationship due to shared antecedents*. For example, it could be that the relationship between being bullied and stuttering is caused by a third factor**.** It is also possible to specify nondirectional relationships between variables, which are sometimes referred to as *unanalyzed prior associations.* In the example in Figure 1, it could be that a person is bullied because they stutter, or that being bullied results in more stuttering. Lastly, *reciprocal effects* may also be accommodated by allowing variables simultaneously to influence each other.

Identification

A model is said to be *identified* when there is a single best value or unique solution for each of its unknown (free) parameters. If a model is not identified it is possible to find an infinite number of values for the parameters that would produce a good fit of the data to the model. *Just-identified* models possess the same number of equations as unknown free parameters. A consequence of this is zero degrees of freedom. *Under-identified* models occur when there are too many parameters to estimate the number of observed measures. The consequences of under-identification may include, impossible parameter estimates, fit tests that are not valid and large standard errors. *Over identified* models are those where there are fewer possible parameter estimates than possible equations. Perhaps the best example of such a model is the well-known multiple regression model. Identified models provide the best evidence in favor of the proposition that the theoretical model represents the data, as there is a unique solution to the data.

Identification is an important concern in SEM, as the methodology provides the end user with the freedom to specify models that are not identified. An SEM is said to be "identified" if the model's restrictions and a population covariance matrix imply unique values for the model's parameters. Assessing identification is complex, given the complexities of matrix algebra. Fortunately, most computer programs provide information in their output if there is a problem regarding identification. Unfortunately, the locus of the problem is not specified.

Estimation

Once a model is specified, the next step is to obtain estimates of the free parameters from the data. To estimate the model based on the data, SEM programs use iterative algorithms. The matrix based on the data, usually the covariance matrix, in SEM is known as **S**. Taking this matrix SEM software calculates a Σ matrix which is the basis of the parameters in the model. SEM software programs compare **S** and Σ iteratively until the similarity between the matrices is maximised.

One of a number algorithms can be used to minimize the similarity between **S** and Σ. These algorithms include Unweighted Least Squares (ULS), Weighted Least Squares (WLS), Generalized Least Squares (GLS), and Maximum Likelihood (ML), which is the most efficient and widely used algorithm of all. Each algorithm has a *fitting function* that generates a value indicating the difference between Σ and **S.** The closer Σ and **S** are, the better the estimate.

The decision about which algorithm to use can be complex. It depends on the nature of the data and the research aims of the study. For example, WLS, GLS, and ML all assume multivariate normality (i.e., the assumption that each variable that is considered is normally distributed, but with respect to each other variable). ULS is unduly affected by the metric of the variables and usually requires prior standardization. This standardization is problematic and should be avoided because standardized covariance (i.e., correlation) matrices result in poorly estimated models (Cudeck, 1989). WLS needs an asymptotic covariance matrix (an estimate of the large sample covariance matrix used to generate weights) and a matrix of correlations to work properly. Current research indicates that the most appropriate estimation method in most cases, including those where the multivariate normality assumption has been violated, is ML (Olsson, Foss, Troye & Howell, 2000).

The estimation of parameter values is achieved by means of an iterative process. This process starts with initial estimates that are sequentially adjusted until model fit cannot be improved. The parameter estimates after iteration constitute the final solution. It is worth noting that, in some cases, the algorithm may fail to explore the entire range of potential values under the fit function and become trapped into a local minimum point. Under such circumstances, the resulting estimates will be suboptimal. In complicated models, it is always sensible to provide the software with alternative sets of starting values to establish whether the resultant solutions converge (this option is not provided by all software). It is also possible for a final solution to contain illogical parameter estimates (e.g., negative variances or correlations over 1.00). These are known as Haywood cases and are indicative of ill-fitting or poorly specified models.

Model fit

Once the estimation process has been completed, it is important to assess how well the data fit the proposed model. A large number of indices have been developed to evaluate the *goodness of fit* of a model and some attention should be given as to which fit indices should be reported. The most widely reported fit index is the χ^2. A significant χ^2 value indicates a poor fit of the model to the data. The rationale for this follows from our earlier discussion of the Σ and **S** matrices. The reader will recall that the closer the matrices, the better the estimate. It follows from this that one does not want a significant difference between the two, as indicated by the χ^2 model fit statistic.

The reason for this is that ideally **S** and Σ will not differ. However, for a test statistic to follow the χ^2 distribution, it is important for the multivariate data to be normally distributed and the sample sizes to be large enough to conform to the requirements of asymptotic distribution theory. Due to its distribution, χ^2 is less likely to be significant in small samples than in large ones. Accordingly, a variety of model fit indices have been developed and these may be categorized according to two basic types, absolute versus incremental (see, e.g., Hu & Bentler, 1995, 1998, 1999).

Two points should be kept in mind in the discussion of fit indices. First, these indices apply to some types of SEM, like the examples in Figure 1, but not to others (e.g., tests of factorial invariance; Cheung & Rensvold, 2002). Second, SEM values parsimony, which can be operationalized as 'the ratio of the degrees of freedom in the model being tested to the degrees of freedom in the null model' (Raykov & Marcoulides, 1999, p. 293). Thus, some fit indices attempt to take parsimony into account by penalizing complex models with many free parameters.

Absolute indices

Absolute fit indices attempt to assess how well the theoretical model reproduces the sample data. The Goodness of Fit Index (GFI) compares the specified model to no model at all. It ranges from 0 to 1, with higher values indicating better fit. The Adjusted GFI (AGFI) introduces an adjustment based on degrees of freedom that penalizes model complexity. It is interpreted in the same way as the GFI. The Standardized Root Mean Square Residual (SRMR) expresses the average discrepancy between observed and expected correlations across all parameter estimates in a model. It ranges from 0 to 1, with lower values indicating better model fit.

A widely used index of absolute fit is the Root Mean Square Error of Approximation (RMSEA), which essentially asks "How well would the model, with unknown but optimally chosen parameter values, fit the population covariance matrix if it were available?" (Browne & Cudeck, 1993, p. 137-138). The RMSEA expresses fit discrepancy per degree of freedom, thus addressing the parsimony of the model. It is generally insensitive to sample size. In contrast to most other indices, it is possible to estimate confidence intervals (usually, 90%) around the point estimate of the RMSEA.

Incremental fit indices

Incremental fit indices measure the proportionate improvement in fit by comparing a target model with a more restricted, nested baseline model (Hu & Bentler, 1999, p.2). Further note, that if two models are equivalent except for a subset of free parameters in model one that are fixed in model two, then model one is said to be a nested model. The Non-Normed Fit Index (NNFI) is an example of such a fit index. Essentially the NNFI compares the specified model with a baseline model (usually an *independence* model, i.e., a model that stipulates that the variables are unrelated. NNFI penalizes model complexity, such that complex models with many free parameters have lower NNFI values.

Bentler (1990) proposed the Comparative Fit Index (CFI), which is based on the non-central χ^2 distribution and ranges from 0 and 1, with higher values indicating better model fit. The CFI penalizes small samples, thereby taking sample size into account.

Modification

Once model fit has been assessed, researchers may wish to adjust their model to account for aspects of the data that do not accord to the theory. This is known as the model modification stage of SEM. It is contentious because it involves freeing up previously fixed parameters on a post-hoc data driven basis. For example, a researcher may fit the model in Figure 1 and subsequently discover that an additional direct path from IQ into severity of stuttering is necessary to improve fit. This modification would result in an improvement of model fit.

Modification indices quantify the expected drop in χ^2 if a previously fixed parameter is set free in order to be estimated from the data and improve the fit of the model. There are three distinct, but asymptotically equivalent, modification indices in SEM, *viz*., the Wald statistic, the Lagrangian Multiplier (LM), and Likelihood Ratio (LR). The Wald test estimates the less restricted model and sees if restrictions should be added. The LM test estimates the more restricted model, and sees if restrictions should be removed. The LR test estimates both models and evaluates the discrepancy in their χ^2 values. The LR test requires two estimates, but accounts for changes in parameter values between the models.

As noted, many are critical of the modification process because it can result in high levels of capitalization on chance. Two strategies are available to avail of the advantages of post-hoc modification strategies without shouldering the attendant pitfalls. First, assuming a sufficient sample size, it is possible to split the data randomly into two sets of approximately equal size. The first data set can be used for post-hoc modifications and model exploration, while the second data set can be used for cross-validation purposes. The second strategy is to collect an independent new dataset and use it to fit the revised model including all the post-hoc modifications.

Modeling strategies

A number of different strategies to conduct SEM have been proposed in the literature. Although all strategies involve the essential steps of model specification, identification, estimation, and modification, the exact manner in which they are carried out can vary very considerably. Perhaps the best known strategy, giving excellent results in psychological studies, is known as the Two Step. Step 1 involves the development of an individual congeneric model for each latent variable. Observed variables which produce goodness of fit in predicting a latent variable are identified. Factor score regression weights (transformed to sum to 1) are used to calculate a composite observed variable from these observed variables. A lower bound estimate of the reliability of the composite observed variable is Cronbach's alpha, but a better measure of reliability can be computed by hand. This is relatively easy to do if there are no error covariances (see Gerbing & Anderson, 1988), but can be difficult if there are (see Werts, Rock, Linn & Joreskog, 1978). Once the reliability is known, Munck (1979) showed that it is possible to fix both the regression coefficients (which reflect the regression of each composite variable on its latent variable) and the measurement error variance. At step 2, the overall model is considered. This involves putting each latent variable and its associated composite observed variable into the whole model. As with most issues there are those who support (Mulaik & Millsap, 2000) and those who refute the use of this method (Hayduk & Glaser, 2000).

Item parceling

In some cases, researchers may be unsure as to how to represent their latent variables. An important choice is between using item parcels versus single items. The sometimes controversial technique of parceling involves summing up a number of individual items in order to construct item parcels and use them as indicators of the latent variables. Critics note that, when variables are not truly unidimensional, parceling may result in model mis-specification and in the acceptance of models that in fact provide a poor fit to the data. On the other hand, MacCallum, Widaman, Zhang and Hong (1999) provide several reasons to use item parcels as indicators of the latent factors. They point out that parceled data are more parsimonious (i.e., there are fewer parameters to be estimated both locally in defining a construct and globally in representing the entire model). Parceled data are also less likely to produce correlated residuals or multiple cross-loadings and may lead to reductions in various sources of sampling error.

In the delinquency literature it has been noted that the use of item parcels overcomes the extreme skewness that is often found in individual items. Thus, a number of authorities in criminology recommend the use of parcels to help achieve multivariate normality (e.g., Farrell & Sullivan, 2000). It is worth noting that the same may apply to stuttering data, particularly in cases of severe skewness. Little, Cunningham, Shahar and Widaman (2002) reviewed the use of parceling in the literature and noted that ‘in the end two clear conclusions may be drawn from our review of the issues. On the one hand, the use of parceling techniques cannot be dismissed out of hand. On the other, the unconsidered use of parceling techniques is never warranted’ (p. 171).

Interactions

SEM is not restricted to linear structural relationships, but can also accommodate interactions and polynomial effects. Indeed, testing interactions through SEM has advantages over the conventional multiple regression approach because SEM overcomes many of the problems that are associated with interaction terms in regression models (e.g., low internal consistency, which reduces the power of the corresponding statistical test). There are several approaches to testing interactions through SEM, their main differentiating characteristic being how the interaction term is represented and estimated (Schumacker, 2002; see also Jaccard & Wan, 1995, Moulder & Algina, 2002).

Criticisms

Critics of SEM methodology (notably Cliff, 1983) have drawn attention to a number of contentious issues and limitations, foremost among which is the erroneous assumption of some practitioners that modeling correlational data can somehow help to establish causal relationships between variables. This criticism perhaps originates from the early use of the term "causal models" to describe SEM analyses.

Another criticism is that researchers often fail to consider alternative models, other than their preferred one, which could provide an equally good or even superior fit to their data. The problem is especially difficult to resolve in cases where a researcher has to grapple with a number of theoretically conflicting, but mathematically equivalent, models (see MacCallum, Wegener, Uchino, & Fabrigar, 1993). Other oft-quoted limitations of SEM *per se* or of the manner in which it is applied include the routine violation of the assumptions on which the analyses are based, the excessive or uncritical reliance on modification procedures, and an unwarranted preoccupation with model fit at the expense of substantive considerations.

Reporting SEMs

A number of texts provide directions on how to report an SEM (e.g., Boomsma, 2000; Hoyle & Panter, 1995). In general the texts suggest the following. A diagram showing the theoretical relations between the elements in the model should be presented in the introduction. This should be much like the one in Figure 1, although it should only include the latent factors. In the results section, prior to conducting the SEM, the strategy adopted (e.g., the Two Step), matrix used (e.g., covariance matrix), and algorithm employed (e.g., MLE) should be reported. The fit indices that should be reported remain an issue of debate. The fit criteria of Hu and Bentler (1999) have become widely adopted in the cases of path analysis and confirmatory factor analysis. It is recommended that a balance be struck between model fit and direct effects. With regard to direct effects, the results should contain all the standardized coefficients and their associated p values. Correlations between error terms should also be reported with an explanation of why they were set to correlate. These values, together with the standardized estimates of the parameters, are reported in a figure, much like the one in Figure 1. Finally, the covariance matrix should be included in the appendix, so that researchers can reproduce the original solutions of authors.

Conclusion

SEM provides a powerful approach to hypothesis testing. It is becoming increasingly popular and, in many cases (even in cognitive psychology where experimentation prevails), it is replacing conventional data analytic techniques, such as exploratory factor analysis and multiple regression. Because SEM encourages researchers to explicitly state their theories and hypotheses in an *a priori* manner, it often leads to more comprehensive and precise theoretical statements. The aim of this short exposition was to bring this flexible and powerful class of modeling methods to the attention of substantive researchers, in the hope that they will prove a useful data analytic tool in the study of the developmental pathways of stuttering. Appendix A takes the reader through a LISREL analysis of manufactured data (these data are presented in Appendix B).

Acknowledgement. The third and the final authors are supported by the Wellcome Trust grant 072639.

References

Baron, R 90. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology,* **51**, 1173-1182.

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin,* **107**, 238 – 246.

Boomsma, A. (2000). Reporting analyses of covariance structures. *Structural Equation Modeling: A Multidisciplinary Journal*, **7**, 461-483.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing Structural Equation Models*. Beverly Hills, CA: Sage

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling,* **9**, 233-255.

Cliff, N. (1983). Some cautions concerning the application of causal modelling methods. *Multivariate Behavioral Research,* **18***,* 115-126.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin,* **105***,* 317-327.

Earley, P. C., & Lituchy, T. R. (1991) Delineating Goal and Efficacy Effects: A Test of Three Models. *Journal of Applied Psychology.* **76** *81-98.*

Farrell, A. D., & Sullivan, T. N. (2000). Structure of the Weinberger Adjustment Inventory Self-Restraint Scale and its relation to problem behaviors in adolescence. *Psychological Assessment,* **12***,* 394 - 401.

Furnham, A., & Davies, S. (2004). Involvement of social factors in stuttering: A review and assessment of current methodology. *Stammering Research,* **1**, 112-122.

Fife-Schaw. C. (2000). Introduction to Structural Equation Modeling. In G. M. Breakwell, S. Hammond & C. Fife-Schaw. (Eds.), *Research Methods in Psychology (Second Edition).* London: Sage.

Gerbing, D. W., & Anderson, J. C. (1988). An updated paradigm for scale development incorporating unidimensionality and its assessment. *Journal of Marketing Research*, **25**, 186-192.

Hayduk, L. A., & Glaser, D. N. (2000). Jiving the four step, waltzing around factor analysis and other serious fun. *Structural Equation Modeling,* **7***,* 1-35.

Hershberger, S. L. (2003). The Growth of Structural Equation Modeling: 1994-2001. *Structural Equation Modeling,* **10**, 35 -47.

Hoyle, R. H. (1995). The Structural Equation Modeling Approach: Basic concepts and fundamental issues. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts Issues and Applications* (pp. 1-18). London: Sage.

Hoyle, R. H., & Panter, (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts Issues and Applications* (pp. 158-176). London: Sage.

Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural Equation Modeling: Concepts Issues and Applications* (pp. 76-99). London: Sage.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods,* **3***,* 424-453.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria vs. new alternatives. *Structural Equation Modeling,* **6***,* 1-55.

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. *Psychological Bulletin,* **116**, 348-357.

Jöreskog, K.G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, **34**, 183-202.

Jöreskog, K.G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, **36**, 409-426.

Jöreskog, K., & Sörbom, D. (1996). *LISREL 8: User's reference guide.* SSI Scientific software International. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kline, R. B. (2004). *Principles and Practice of Structural Equation Modeling (Second Edition).* NY: Guilford Press.

Kolk, H., & Postma, A. (1997). Stuttering as a covert repair phenomenon. In R. F. Curlee & G. Siegel, (Eds.). *Nature and treatments of stuttering: New directions*. Needham Heights, MA: Allyn & Bacon.

Levine, S. Z., & Jackson, C. J. (2004). Eysenck's theory of crime revisited: Factors or primary scales? *Legal and Criminological Psychology,* **9**, 135-152.

Little, T. D., Cunningham, Shahar, G., Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling,* **9***,* 151-173.

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin,* **114**, 185-199

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods,* **4**, 84-99.

Maruyama, G. M. (1998). *.Basics of Structural Equation Modeling*. Thousand Oaks: Sage.

Moulder, B. C., & Algina, J. (2002). Comparison of methods for estimating and testing latent variable interactions. *Structural Equation Modeling,* **9**, 1-19.

Mulaik, S. A., & Millsap, R. E. (2000). Doing the four step right. *Structural Equation Modeling,* **7***,* 36-73.

Munck, I. M. (1979). Model building in comparative education : Application of the LISREL method to cross-national survey data. *International Association for the Evaluation of Educational Achievement Monograph Series*, No. 10. Stockholm: Almqvist & Wiksell.

Muthén, B. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, **29**, 81-117.

Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling,* **7**, 557-595.

Quintana, S. M., & Maxell, S. E. (1999). Implications of recent developments in structural equation modeling for counseling psychology. *The Counseling Psychologist,* **27**, 485-587.

Petrides, K. V., Furnham, A., Jackson, C. J., & Levine, S. Z. (2003). Exploring issues of personality measurement and structure through the development of a revised short form of the Eysenck Personality Profiler. *Journal of Personality Assessment*, **81**, 272-281.

Raykov, T., & Marcoulides, G.A. (1999). On desirability of parsimony in structural equation model selection. *Structural Equation Modeling,* **6**, 292-300.

Raykov, T., & Marcoulides, G. A. (2001). Can there be infinitely many models equivalent to a given covariance structure model? *Structural Equation Modeling,* **8**, 142-149.

Riley, G. D. (1994). *The Stuttering Severity Instrument for Adults and Children* (third edition). Austin, TX: Pro-Ed.

Schumacker, R. E. (2002). Latent variable interaction modeling. *Structural Equation Modeling,* **9**, 40-54.

Smith, A. (1999). Stuttering: A unified approach to a multifactorial dynamic disorder. In N. Bernstein Ratner & E. C. Healey (Eds.), *Stuttering research and practice: Bridging the gap.* Mahwah: Lawrence Erlbaum Associates.

Smith, A., & Kleinow, J. (2000). Kinematic correlates of speaking rate changes in stuttering and normally fluent adults. *Journal of Speech, Language, Hearing Research*, **43**, 521-536.

Werts, C. E., Rock, D. R., Linn, R. L., & Jöreskog, K. G. (1978). A general method of estimating the reliability of a composite. *Educational and Psychological Measurement*, **38**, 933-938.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics,* **5**, 161-215.

Appendix A

This Appendix takes readers through the steps involved in a LISREL 8.3 analysis (software published by Scientific Software International, Inc.). The problem for analysis was based on the IQ/bullying examples used in the body of the text using model one as a basis (it was not possible to test the other model in the paper as it did not converge.) As shown in Figure 1, model one has three latent variables (IQ, stuttering severity, SS and bullying BU) which correspond to ach (achievement), ss (stuttering severity) and vic (victimization) in the current analysis. According to model one, IQ is reflected in vo (vocabulary) si (similarities) and co (comprehension) test scores and the equivalent here is that achievement is reflected in IQ and GPA scores. In model one, overall stuttering severity (ss) is reflected in self (Srs), peer (Prs) and teacher's (Trs) reports of severity of stuttering, and here SS is reflected in parent, child and clinical (PARENT, CHILD SSITOTAL, respectively) scores. The final latent variable of bullying (Bu) in model one is reflected in self (srb) peer (prb) and teacher (trb) reports of bullying and the measure of bullying here (vic) is reflected in peer (VICTIM), parent (VICTIM2) and teacher's assessments of bullying (SBS).

The data are made up for this exercise and, therefore, the conclusions from the analysis have no meaning. The data are given in Appendix B so that readers can reproduce the analysis if they wish. The data contain the following observed variables (labels are those used in the LISREL analysis that follows).

SSITOTAL	- clinical assessment of stuttering severity - high scores = more severe
PARENT	- parent report of stuttering severity - high scores = more severe
CHILD	- child report of stuttering severity - high scores = more severe
IQ	- score on the Otis-Lennon Mental Abilities Test
GPA	- grade point average
SBS	- teacher report of bullying - low scores = bullied a lot
VICTIM	- peer reports of bullying - high scores = more nominations as bully victim
VICTIM2	- parent report of bullying - high scores = bullied a lot

The following is the SIMPLIS syntax used to produce the SEM in LISREL.

```
Title Testing a Path Model of Stuttering
 Observed Variables:   SSITOTAL PARENT CHILD IQ  GPA SBS
VICTIM VICTIM2
 Covariance Matrix from File Made.cov
Sample Size: 112
Latent Variables:
 ach ss vic
Relationships:
 ach -> IQ GPA
 ss -> PARENT CHILD SSITOTAL
 vic -> VICTIM VICTIM2 SBS
 vic = ach
 ss = vic
 ss = ach
Path Diagram
End of Problem
```

A description of each of the steps in the SIMPLIS code now follows. The code itself is given in Courier font and the comments are in Times New Roman:

1. `Title Testing a Path Model of Stuttering` – the keyword 'TITLE' gives a name to the analysis for later reference.
2. `Observed Variables:   SSITOTAL PARENT CHILD  IQ  GPA SBS VICTIM VICTIM2` – The command 'Observed variables' tells the program that the labels that follow are

the observed variables. For instance, the IQ measures are scores on the Otis-Lennon Mental Abilities Test

3. `Covariance Matrix from File Made.cov -` This is the basic covariance matrix which may be obtained by a package such as SPSS. The covariance matrix has been saved in file Made.cov.
4. `Sample Size: 112 –` This tells the computer the number of observation in the sample. A sample size of 112 is small for such analyses. It is difficult to give concrete advice as to what sample size is appropriate. One way is to take 200 as an absolute minimum, but it should also be ensured that the ratio of variables to factors is as high as possible. Naturally, like more conventional forms of analysis Power Analysis may serve as a useful guide.
5. `Latent Variables: ach ss vic –` This line identifies which are the latent variables. There are three here, achievement (ach), severity of stuttering (ss) and victimization (vic).
6. `ss -> PARENT CHILD SSITOTAL –` This line indicates that the latent variable severity of stuttering 'ss', causes PARENT, CHILD and SSITOTAL (parent, child and clinical assessment of severity of stuttering, respectively).
7. `vic -> VICTIM VICTIM2 SBS` This line indicates that the latent variable 'vic' (victimization), causes VICTIM VICTIM2 SBS (peer, parent and teacher indications about bullying).
8. `vic = ach` This line indicates that the latent variable 'vic' (victim of bullying), causes the latent variable ach.
9. `ss = vic -` This line indicates that the latent variable 'ss' (severity of stuttering) causes the latent variable 'vic'.
10. `ss = ach -` This line tells the program that latent severity of stuttering is caused by achievement
11. `Path Diagram` This line tells the program to produce a path diagram.
12. `End of Problem` This line tells the program that it is the end of the problem.

LISREL then produces the following output (comments that have been added for explanation are given in bold Times New Roman font):

```
                              DATE: 12/20/2004
                                TIME: 18:07
                              L I S R E L  8.30
                                     BY
                       Karl G. Jôreskog & Dag Sôrbom
                  This program is published exclusively by
                  Scientific Software International, Inc.
                     7383 N. Lincoln Avenue, Suite 100
                      Lincolnwood, IL 60712, U.S.A.
            Phone: (800)247-6113, (847)675-0720, Fax: (847)675-2140
       Copyright by Scientific Software International, Inc., 1981-2000
           Use of this program is subject to the terms specified in the
                        Universal Copyright Convention.
                          Website: www.ssicentral.com
The following lines were read from file C:\LISREL83\DATA\STUT.SPL:
Title Testing a Path Model of Stuttering
Observed Variables:   SSITOTAL PARENT CHILD  IQ  GPA SBS VICTIM VICTIM2
Covariance Matrix from File Made.cov
Sample Size: 112
Latent Variables:
 ach ss vic
Relationships:
ach -> IQ GPA
 ss -> PARENT CHILD SSITOTAL
 vic -> VICTIM VICTIM2 SBS

 vic = ach
 ss = vic
 ss = ach
```

```
Path Diagram
 End of Problem
 Sample Size =   112

Testing a Path Model of Stuttering

        Covariance Matrix to be Analyzed

           SSITOTAL     PARENT      CHILD        SBS     VICTIM    VICTIM2
           --------   --------   --------   --------   --------   --------
 SSITOTAL     68.65
   PARENT     38.99      40.27
    CHILD     29.91      21.19      27.62
      SBS    -21.73     -13.11      -8.61      14.75
   VICTIM     15.27       8.77       6.26      -5.16       6.46
  VICTIM2      6.00       3.94       2.55      -2.80       1.66       1.38
       IQ    -83.65     -58.84     -42.09      27.95     -17.62      -7.81
      GPA     -4.95      -3.66      -2.47       1.67      -1.02      -0.43

        Covariance Matrix to be Analyzed

                 IQ        GPA
           --------   --------
       IQ    170.13
      GPA      7.83       0.70

Testing a Path Model of Stuttering
 Number of Iterations = 35
 LISREL Estimates (Maximum Likelihood)
         Measurement Equations
```

Below are the measurement equations

```
 SSITOTAL = 7.77*ss, Errorvar.= 8.32 , R² = 0.88
                               (2.19)
                                3.81

   PARENT = 5.15*ss, Errorvar.= 13.72, R² = 0.66
           (0.42)              (2.08)
            12.38               6.61

    CHILD = 3.76*ss, Errorvar.= 13.45, R² = 0.51
           (0.39)              (1.92)
            9.70                7.00

      SBS = 2.90*vic, Errorvar.= 6.34 , R² = 0.57
                                (1.05)
                                 6.04

   VICTIM =  - 1.96*vic, Errorvar.= 2.62 , R² = 0.59
              (0.24)               (0.45)
              -8.01                 5.87

  VICTIM2 =  - 0.85*vic, Errorvar.= 0.65 , R² = 0.53
              (0.11)               (0.10)
              -7.54                 6.27


       IQ = 11.44*ach, Errorvar.= 39.16, R² = 0.77
           (1.02)                (8.69)
            11.25                 4.51

      GPA = 0.68*ach, Errorvar.= 0.23  , R² = 0.67
           (0.067)              (0.040)
            10.17                5.84
```

These are the relationships between the latent variables.

```
         Structural Equations

   ss =  - 0.46*vic - 0.59*ach, Errorvar.= 0.0098, R² = 0.99
          (0.12)     (0.12)               (0.037)
          -3.78      -5.01                 0.26

  vic = 0.80*ach, Errorvar.= 0.36 , R² = 0.64
       (0.11)                (0.12)
        6.97                  3.12


         Reduced Form Equations

   ss =  - 0.96*ach, Errorvar.= 0.088, R² = 0.91
          (0.082)
          -11.60

  vic = 0.80*ach, Errorvar.= 0.36, R² = 0.64
       (0.11)
        6.97

         Correlation Matrix of Independent Variables

                 ach
            --------
                1.00


         Covariance Matrix of Latent Variables

                  ss        vic        ach
            --------   --------   --------
       ss       1.00
      vic      -0.93       1.00
      ach      -0.96       0.80       1.00
```

These are the 'Goodness of Fit Indices'

```
                          Goodness of Fit Statistics

                            Degrees of Freedom = 17
              Minimum Fit Function Chi-Square = 25.31 (P = 0.088)
     Normal Theory Weighted Least Squares Chi-Square = 24.55 (P = 0.11)
                Estimated Non-centrality Parameter (NCP) = 7.55
           90 Percent Confidence Interval for NCP = (0.0 ; 24.86)

                       Minimum Fit Function Value = 0.23
               Population Discrepancy Function Value (F0) = 0.068
              90 Percent Confidence Interval for F0 = (0.0 ; 0.22)
            Root Mean Square Error of Approximation (RMSEA) = 0.063
            90 Percent Confidence Interval for RMSEA = (0.0 ; 0.11)
              P-Value for Test of Close Fit (RMSEA < 0.05) = 0.32

                Expected Cross-Validation Index (ECVI) = 0.56
            90 Percent Confidence Interval for ECVI = (0.50 ; 0.72)
                        ECVI for Saturated Model = 0.65
                       ECVI for Independence Model = 5.69

   Chi-Square for Independence Model with 28 Degrees of Freedom = 615.41
                          Independence AIC = 631.41
                              Model AIC = 62.55
                            Saturated AIC = 72.00
                          Independence CAIC = 661.16
                             Model CAIC = 133.20
                           Saturated CAIC = 205.87
```

```
                 Normed Fit Index (NFI) = 0.96
             Non-Normed Fit Index (NNFI) = 0.98
         Parsimony Normed Fit Index (PNFI) = 0.58
             Comparative Fit Index (CFI) = 0.99
             Incremental Fit Index (IFI) = 0.99
               Relative Fit Index (RFI) = 0.93

                    Critical N (CN) = 147.56


            Root Mean Square Residual (RMR) = 0.78
                    Standardized RMR = 0.033
             Goodness of Fit Index (GFI) = 0.95
        Adjusted Goodness of Fit Index (AGFI) = 0.89
      Parsimony Goodness of Fit Index (PGFI) = 0.45
```

These are the suggested parameters to add, i.e., the modification indices.

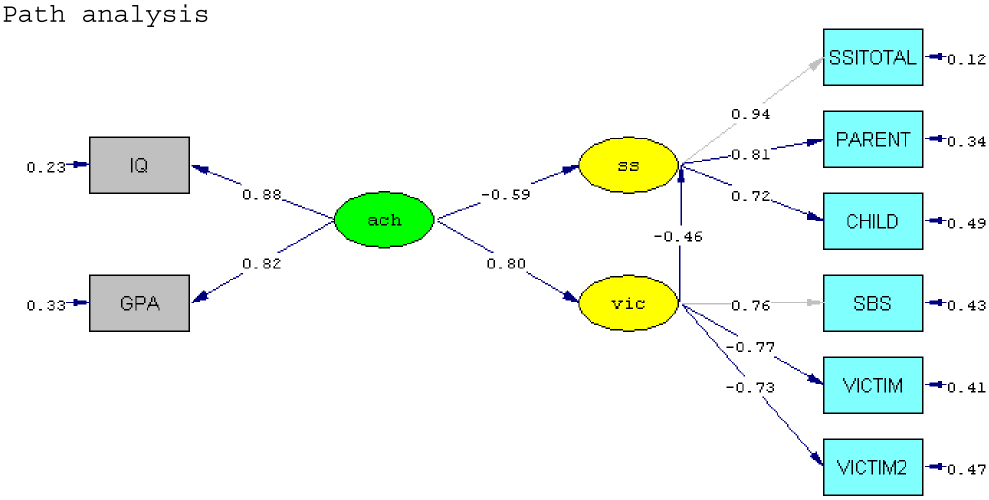
```
       The Modification Indices Suggest to Add the
 Path to   from      Decrease in Chi-Square    New Estimate
SSITOTAL   vic                 7.9                 -5.85

The Modification Indices Suggest to Add an Error Covariance
 Between    and      Decrease in Chi-Square    New Estimate
VICTIM    SSITOTAL             9.6                  2.30


      The Problem used    12760 Bytes (=  0.0% of Available Workspace)

                          Time used:    0.047 Seconds
```

Path analysis



Chi-Square=24.55, df=17, P-value=0.10523, RMSEA=0.063

The path diagram with standardized estimates is produced in the final step in the analysis. Note that all direct effects are significant with the exception of SS to SSITOTAL and the latent variable relating vic to SBS. The latent variable SS reflects estimates of child and parent indications about severity of the problem, but not clinical indications about severity of stuttering. The model indicates that both achievement and victimization influence severity of stuttering, and that achievement has both a direct effect on severity of stuttering and an indirect effect that is mediated by victimization. This model was estimated using a covariance matrix with maximum likelihood estimation. It appeared to fit the data relatively well (according to the criteria of Hu & Bentler, 1999) although it is limited by the small sample size ($\chi^2 = 24.55$, df=17, P =0.11, RMSEA=0.06; Standardized RMR = 0.03, CFI = 0.99, NNFI = 0.98).

Appendix B

Fictitious data compiled for modeling purposes only. These are included so that readers may reproduce the results or practice developing different forms of SEM analysis.
SSI = clinical assessment of stuttering severity - high scores = more severe; *Parent* = parent report of stuttering severity - high scores = more severe; *Child* = child report of stuttering severity - high scores = more severe; *IQ* = score on the Otis-Lennon Mental Abilities Test; *GPA* = grade point average; *SBS (social behavior at school checklist)* = teacher report of bullying - low scores = more bullying; *Victim1* = peer reports of bullying - high scores = more nominations as bully victim; *Victim2* = parent report of bullying - high score = more bullying.

SSI	Parent	Child	IQ	GPA	SBS	Victim1	Victim2
30	27	28	79	1.67	29	1	2
23	21	20	108	4.00	38	1	1
14	10	11	131	3.75	42	0	1
25	10	25	100	2.50	40	2	1
18	20	19	107	3.50	41	1	1
12	13	17	115	4.00	42	0	1
15	14	19	92	2.23	42	1	2
8	10	8	111	3.00	41	1	1
9	17	17	95	3.00	40	0	1
6	14	13	106	3.75	41	0	1
9	9	9	107	3.00	42	1	2
29	25	22	88	2.25	38	3	3
19	18	15	109	2.25	42	1	3
19	14	8	102	2.75	40	2	1
25	20	34	81	2.00	39	3	2
36	23	20	93	2.00	35	7	2
19	17	18	106	2.75	38	1	1
30	20	20	97	2.67	25	4	5
21	10	20	112	3.00	42	2	2
12	15	15	95	2.75	40	0	1
29	39	33	86	1.00	33	1	2
24	34	22	90	2.50	40	0	3
15	18	13	118	3.00	42	1	2
13	20	21	115	3.75	42	0	1
11	19	16	105	2.75	40	0	1
22	26	18	81	1.50	41	5	2
35	35	26	83	.67	30	6	3
27	22	29	85	1.75	40	1	3
24	19	25	83	2.75	38	0	5
20	19	16	105	2.50	39	1	1
8	18	20	115	4.00	42	1	2
29	24	30	102	2.75	42	3	3
28	32	17	81	1.50	40	3	2
16	13	15	118	3.00	40	1	2
9	17	15	120	4.00	42	0	1
8	13	11	129	3.75	41	1	2
11	19	9	111	3.00	40	0	2
19	20	10	100	2.50	35	0	2
11	17	22	109	2.25	40	1	1
25	25	29	81	2.00	39	1	1
12	20	17	112	3.00	42	1	2

31	22	18	93	2.00	40	2	1
29	30	22	86	1.00	36	5	3
30	26	24	85	2.50	32	5	3
18	25	12	97	2.67	42	1	1
37	23	22	85	1.75	31	7	2
28	26	22	79	1.67	35	3	4
20	19	15	95	3.00	42	2	4
33	26	24	90	2.50	40	7	2
34	27	24	83	.67	39	8	2
34	25	22	92	2.23	38	4	3
16	14	19	115	3.75	42	2	2
28	26	22	88	2.25	40	2	2
18	10	13	106	2.75	42	2	1
10	15	13	108	4.00	41	0	1
11	14	11	115	3.75	41	0	2
17	19	16	106	3.75	40	2	3
25	27	23	124	2.00	37	6	3
24	27	14	95	2.75	36	2	3
25	15	13	85	2.75	32	2	4
22	19	21	99	3.50	42	1	2
28	24	21	99	2.67	40	8	3
28	22	21	95	2.75	36	4	3
25	24	18	92	3.50	39	2	2
36	32	25	83	2.75	30	9	5
22	20	19	95	1.50	40	2	2
25	23	21	92	2.23	39	4	3
23	21	15	105	2.75	35	1	2
30	24	23	85	1.50	34	2	3
8	15	16	115	4.00	42	0	1
14	17	15	111	3.00	42	4	1
18	16	18	112	3.00	41	1	2
11	15	17	118	3.00	42	0	2
23	18	14	100	2.50	36	0	3
25	32	21	95	1.50	36	2	3
27	22	23	88	2.25	31	1	2
24	27	26	86	1.00	42	0	2
29	30	22	85	1.75	40	4	3
10	13	17	115	3.75	42	1	1
8	10	12	129	3.75	41	0	2
15	22	11	106	2.75	40	2	1
31	24	21	90	2.50	30	10	4
15	25	23	99	3.00	41	4	2
34	28	23	81	2.00	31	9	5
32	35	25	79	1.67	33	6	4
30	28	24	93	2.00	40	5	3
18	19	18	109	2.25	42	2	2
26	25	23	83	2.75	40	4	3
28	26	23	85	2.50	38	6	2
26	25	17	85	2.75	40	3	3
33	30	25	83	.67	35	5	4
19	20	20	105	2.75	42	4	1
19	16	18	115	3.00	40	0	2
18	20	19	115	3.00	39	1	1

31	29	25	83	.67	31	6	6
9	15	18	115	4.00	42	1	1
25	22	22	99	3.00	40	2	2
11	19	15	115	4.00	42	0	2
12	15	16	115	3.75	42	0	3
24	21	21	99	2.75	41	4	2
29	26	22	88	2.25	40	5	3
23	20	19	109	2.25	38	2	2
26	22	22	100	2.50	35	0	3
40	32	29	85	1.75	29	10	6
30	25	22	93	2.00	37	6	5
31	22	22	90	2.50	36	5	2
20	15	18	118	3.00	41	0	1
30	33	23	95	1.50	35	0	4
29	22	23	92	3.50	38	5	3
28	30	21	92	2.23	38	4	3
35	29	30	86	1.00	30	8	4
25	19	22	97	2.67	40	2	2

The effect of using time intervals of different length on judgements about stuttering

Peter Howell
Department of Psychology, Centre for Human Communications, Institute of Cognitive Neuroscience, and Institute of Movement Neuroscience, University College London, Gower St., London WC1E 6BT
p.howell@ucl.ac.uk

Abstract. Conventional clinical procedures for assessment of stuttering are reported to have poor reliability. Time interval analysis procedures have been reported to produce greater reliability than the conventional procedures. In time interval procedures, successive intervals of the same duration are extracted from a sample of speech and judged by participants as stuttered or fluent. There is a problem insofar as the amount of speech judged stuttered depends on the length of the interval used. This problem is illustrated in an experiment in which 1-s and 5-s intervals were drawn from the same samples of speech and judged by participants as stuttered or fluent. It is also shown that the problem of lack of sensitivity when longer intervals are used is more acute for individuals who exhibit severe stuttering. Since ability to detect changes in stuttering rate is dependent on the length of interval used (as well as stuttering severity), the procedure can highlight or disguise changes in stuttering rate depending on parameterization of interval length and choice of participants to study. Thus, use of different length intervals across studies can distort whether particular treatments have an effect on speech control. Therefore, it is concluded that time interval analysis, as it is currently used, is an unsatisfactory procedure. If a standard-length interval could be agreed, comparison across studies or analyses would be possible. **Keywords:** Stuttering assessment, time interval analysis procedure.

1. Introduction

A conventional method for assessing stuttering is to make a count of the disfluent events directly from recordings (by using, for instance, a manually operated counter). Studies have shown that agreement between judges using such methods is only around 60% (e.g., Curlee, 1981; Martin & Haroldson, 1981). Ingham and his colleagues have proposed that time-interval (TI) procedures produce higher agreement. In TI procedures applied to stuttering, listeners hear (and in some cases see) a fixed-length extract of speech and designate it as stuttered (STUT) or fluent (FLU). The Ingham group claim that such procedures ".. could certainly lead to different ways of overcoming the problem of judgement reliability for stuttering" (Ingham, Cordes & Gow, 1993, p.512). If this claim is true, then TI procedures should be used for clinical assessment in preference to the standard procedures. The current study examines the claim that TI procedures provide a reliable indication about stuttering.

Reliability can be defined as the relationship between observed and true scores (Cordes, 1994). Ingham and co-workers used expert judges to establish the true responses to each time interval. They employed audio-visual recordings of speech from a selected sample of speakers who stutter. The samples of speech from each speaker were divided up into adjacent, non-overlapping 4-s intervals and these were played to judges in semi-random order. Each interval was presented for judgement to the panel of four experienced judges who independently assessed each sample twice. For each 4-s sample, these judges were instructed to identify whether the sample contained stuttering or not (Ingham et al., 1993, p.506). These data were then used to locate which intervals the experienced judges agreed on (a criterion of 7/8 judgements given the same response was used for this purpose) and to establish what the agreed response should be (the response given on the majority of occasions for the agreed intervals). 110 of the 143 intervals tested (77%) were agreed by the experienced judges and 64 of these were judged to be STUT.

A major potential source of bias in such procedures stems from the fact that intervals of different lengths have been used across different TI studies. Howell, Staveley, Sackin and Rustin (1998) pointed out that a) when long intervals are used, the chance of the interval containing signs of stuttering increases, and b) long intervals will tend to result in a ceiling effect with all intervals judged as being stuttered. These effects limit the use of TI procedures for assessing whether a treatment reduces stuttering rate. To make such an assessment, a researcher may decide to have participants' speech judged before and after treatment. If an interval-length is used which results in a ceiling effect before and after treatment, it would not be possible to detect a change due to the treatment. Changes due to the treatment may have been evident if a shorter interval had been used. A worked example shows this. Say there are two samples of speech, each 100 s in length. One contains 30 stutterings and the other 20 stutterings and in both cases these are evenly distributed over the sample[1]. Assuming a speech rate of five syllables per second (Perkins, 2001) this would correspond to stuttering rates of 6%, and 4%, respectively. If the samples are partitioned into 5-s TIs, there will be an average of one stutter for the 20 stutter sample and 1.5 for the 30 stutter sample. If these are reliably judged, 100% of intervals will be designated STUT for both samples. This would suggest that there is no difference in stuttering rate between the samples. Different outcomes would be expected if the same samples were partitioned into 1-s TIs. The 1-s TIs made from the first sample would include 30 stuttered intervals and the second sample 20 stuttered intervals and observers making accurate judgements would reflect a 10% difference in stuttering rate between the two sets of intervals. The ceiling effect when the longer intervals are used shows that these intervals may fail to detect any fluency-enhancing effects of treatment procedures, which would have been evident if shorter intervals had been used. There are indications that this is more than a hypothetical possibility. For instance, Ingham, Moglia, Frank, Costello-Ingham and Cordes (1997) assessed the effects of frequency-shifted feedback on the speech of people who stutter using 5-s TIs. The long interval appears to have a) resulted in a ceiling effect in both the pre- and post-treatment conditions (both conditions showed around 100% intervals judged to be STUT), and b) led in turn to failure to find effects of the treatment that the majority of participants reported (i.e., increased fluency under frequency-shifted feedback) (Howell et al., 1998).

This analysis shows that the authors might very well have found fluency-enhancing effects of frequency-shifted feedback, in line with what their participants reported, if they had used a shorter (more sensitive) TI. Clearly, a procedure that can result in misleading conclusions about treatment outcome is not satisfactory. To provide evidence on this, the study compared STUT/FLU judgements of the same material when it was segmented into 5-s and 1-s TIs. The earlier analysis predicts that more speech intervals will be judged STUT when the longer (5-s) intervals are used than when the shorter (1-s) intervals are used. The influence of interval length on speakers with different severity of stuttering is also examined. The above analysis predicts that speakers with more severe stutters will be less affected (as a higher proportion of their intervals will be judged stuttered whatever the length of interval used).

2. Method

Speech samples. Eight 2-minute samples of speech from the UCLASS archive of stuttered speech (Howell & Huckvale, 2004) were selected for use in this study. These represent one sample each from eight male speakers ranging in age from 7 years 7 months to 17 years 9 months. The samples were chosen to cover a broad range of both age and stuttering rate. The samples can be down-loaded and examined by visiting http://speech1.psychol.ucl.ac.uk/index.htm. The samples employed were 0030_17y9m.1, 0061_14y8m.1, 0078_16y5m.1, 0095_7y7m.1, 0098_10y6m.1, 0138_13y3m.1, 0210_11y3m.1, 0234_9y9m.1 (as listed in the first column of Table 1). Further details about these (and other) speech samples from the UCLASS archive can be obtained from Howell and Huckvale (2004).

Procedure. A MATLAB program was written to divide a file into either 1-s or 5-s intervals. The MATLAB program played one set of intervals and recorded listeners' responses to each interval as STUT or FLU. The program had an option that allowed intervals to be replayed at random with no replacement, or in sequence. The responses of the experts were used to obtain intervals that were agreed as fluent (FLU) or stuttered (STUT)[2]. These provided criteria against which the responses of naïve judges were assessed. These are described in more detail below.

Selection of intervals for assessment and experts' criteria judgements. Four expert judges, each of whom had 10 or more years' intensive experience of assessing stuttered speech, were employed. The procedure for assessing the intervals was broadly similar to that which Ingham et al. (1993) adopted (see the start of the introduction). The differences in procedure were 1) that the experts judged each sample of speech both when it was segmented into 1-s, and into 5-s, TIs, 2) that judgements about the 1-s, and 5-s TIs were made twice, once when the TIs were sampled at random without replacement, and once when the TIs were presented in sequential order (the reasons for the latter are given below), and 3) they were instructed as to what events should be used to judge an interval as STUT, (this ensured that the task was less open-ended than in other studies).

The reason that judgements were made in sequential and random order was that in initial work with the experts, they indicated that their designations might well have differed if they had had the preceding interval available as context. (Howell, Sackin & Glenn, 1997 also found that judgements were better if some of the prior context, the preceding word in their report, was given before the judged extract.) The segmentation process led to two artifacts: 1. Sometimes an interval started at a point in speech that gave the sound an apparent hard onset which led judges to rate it as STUT. 2. Truncation at the end of an interval sometimes made an interval sound as if the first sound was part of a repetition sequence. In both cases the sound would have been designated STUT while hearing the preceding context would have led the judge to revise their view and designate the interval as FLU. There is a five times greater chance of these artifacts affecting 1-s than 5-s intervals. A 5-s TI was dropped when the judgement about one or more of the five 1-s intervals that made up the 5-s interval was judged STUT while the 5-s interval itself was judged FLU (indicative of these problems) in more than one of the eight judgements made by the experts.

It is also possible that there are cases where stutterings were not apparent on 1-s intervals but were on 5-s intervals. An example would be where a prolonged phone is split between two adjacent 1-s TIs leaving each section of the prolonged phone below the duration-threshold of a prolongation that then leads to each interval judged FLU. A 5-s TI was dropped when the judgement about the 5-s interval was STUT while all the five 1-s intervals it contained were judged FLU (indicative of such a problem) for more than one of the eight judgements made by the experts.

To ensure consistency in what events were considered as STUT, the experts were told:

1.) not to count revisions, whole word repetitions or retraces as STUT unless there were other signs of stuttering (e.g., prolongation or repetition of part of a word).
2.) to count silence, laughs, breathing noises and filled pauses as FLU. Intervals with these events were included in the results of the naïve judges who were also told to judge them as fluent.

The experts were also allowed to indicate any intervals that they considered to be ambiguous with repect to STUT/FLU status. One situation where this arose was when the duration of a sound was prolonged only marginally as this could have been done for emphasis or might have been a brief period of disfluency.

Ingham et al.'s (1993) procedure depended on the experts using their own judgement as to what was, or was not, stuttered. This would inevitably lead to difference of opinion and, for this reason, the procedure employed here where judges were told what events to consider STUT was considered preferable.

The experts next indicated when there were extraneous sounds, as these might have affected productions and/or judgements in intervals that contained them. These usually arose because the participant knocked the microphone or some other object in the recording environment. There were also occasional prompt questions by the researcher making the recordings (when, for instance, the speaker ran out of things to say). These intervals were noted and intervals with these events were not included in the results of the naïve judges.

During any one session, judgements were made about either the 1-s or 5-s set of intervals (interval length) and the intervals were either presented in random, or sequential, order (order type). The speech of each person who stutters was assessed in turn. The interval length by order type judgements were carried out in different random orders by the expert judges, and the judgements were separated by at least a week to prevent carry-over effects on judgements. The experts indicated which rejection criteron applied to an excluded interval. The numbers of 5-s intervals excluded by the different criteria discussed above are summarised in Table 1. 86 5-s intervals were excluded in total (corresponding to 430 1-s intervals).

Table 1. Summary of number of 5-s intervals that were excluded (under the column labeled N) and the percentage of total speech available this represents (under the column labeled %) for the exclusion criteria indicated at the left of each row.

	N	%
Artefacts of segmentation process:		
5-s FLU, one of the five 1s-intervals agreed STUT	24	12
5-s STUT, all of the five 1s-intervals agreed FLU	10	5
Ambiguous	28	14
Extraneous noises	24	12
Total	86	43

Note that there were more cases where the 5-s interval was judged FLU but one of the constituent 1-s intervals was judged STUT (24) than vice versa (10). This suggests that the hard onset and spurious repetition artifacts were more prevalent than cases where the more extensive 5-s context allowed disfluencies to be detected that were missed when the shorter 1-s segments were used.

Some of the 1-s intervals that were agreed by the experts occurred within a 5-s interval that included other 1-s intervals that the experts did not agree on (six cases in total). All constituent 1-s intervals of a 5-s interval were dropped when one or more of the 1-s intervals were not agreed. This permitted direct comparison – i.e. all 1-s intervals and the 5-s interval they comprised were agreed.

The exclusion criteria affected speakers differentially and this depended on the severity of their disorder. Table 2 gives some details of how the exclusion criteria affected individual speakers and how this relates to stuttering incidence in the 5-s intervals that remain. The overall rejection rate of 5-s samples was 46% (i.e., the 43% in Table 1 and the additional 3% where 5-s agreed intervals contained at least one 1-s interval not agreed on). The same applies (though to a lesser extent) to Ingham et al. (1993). As the main point here is to evaluate the TI procedure for intervals that are precisely defined, more strict criteria were applied (leading to the higher rejection rate than in Ingham et al., 1993).

Table 2. Speakers are labeled at the top of the Table. The first row gives the duration (in s) of the original file (based on 5-s intervals) and results for across speakers (labeled N). The second row indicates the duration (again in s) after all exclusion criteria were applied. The amount of speech lost (in s) is given in the third row (i.e. the difference between what was available initially and after the exclusion criteria were applied) and row four gives this as a percentage of the total material available. Row five gives the TIs that were agreed to be STUT (in s) for the data after the exclusion criteria were applied and row six represents this as a percentage of all material (row two).

	Speaker								
	1	2	3	4	5	6	7	8	N
Initial duration	125	130	125	125	120	130	125	120	1000
After exclusion	55	105	75	75	55	40	85	50	540
Lost	70	25	50	50	65	90	40	70	460
%age lost	56	19	40	40	54	69	32	58	46
TIs where judged STUT	40	105	45	75	5	20	70	45	405
%age stuttered	73	100	60	100	9	50	82	90	75

The experts' judgements served two roles: 1) to select intervals for testing with the naïve judges as described in the next section; 2) to determine whether the selected intervals were responded to correctly by naïve judges (correspond with experts' responses) or not (did not correspond with experts' responses).

Assessment of intervals by naïve judges. The naïve judges were undergraduates aged betwen 20 and 22 from a variety of humanity disciplines who reported that they had no experience of judging stuttered speech (speech science students were explicitly excluded). Eight naïve judges assessed the sets for each interval length (1-s and 5-s intervals) in random order for each speaker separately in the same way as the experts. The two assessments of the same material (1-s or 5-s intervals) were done at least a week apart. TI judgements were made with intervals presented as with the experts. All intervals were judged (i.e., material in the row labeled 'initial duration' in Table 2) but the results are only reported for those intervals that the experts agreed on (i.e., material in the row labeled 'after exclusion' in Table 2). Using all material ensured that there was at least one FLU interval agreed by the experts for each of the speakers, so judges should have used both available responses and the context in which they made these judgements was the same as that of the experts so that judgements could be compared (Parducci, 1965).

3. Results

Prediction one

The experts' response designations (STUT or FLU) for the agreed intervals, were used as the criterion against which to assess the accuracy of the naïve judges. The expert judgements for 1-s intervals excluded all five 1-s intervals when the 5-s interval they comprise was agreed by the experts to be FLU but one of its constituent 1-s TI was judged STUT and also excluded all five 1-s intervals when the 5-s interval they comprise was agreed by the experts to be STUT but all the constituent 1-s TI were judged FLU. Cases were, however, included where 5-s intervals were agreed to be STUT by the expert judges which contained one or more 1-s intervals that were agreed by the experts to be FLU (though at least one 1-s interval has to be expert-agreed STUT). This arises when the experts agree that there is a stuttering of less than 5-s in length and leads to agreed FLU 1s-TI from the expert judges for where there were no agreed FLU 5-s TIs for speakers 2 and 4 (see below where responses of each judge to each speaker are presented).

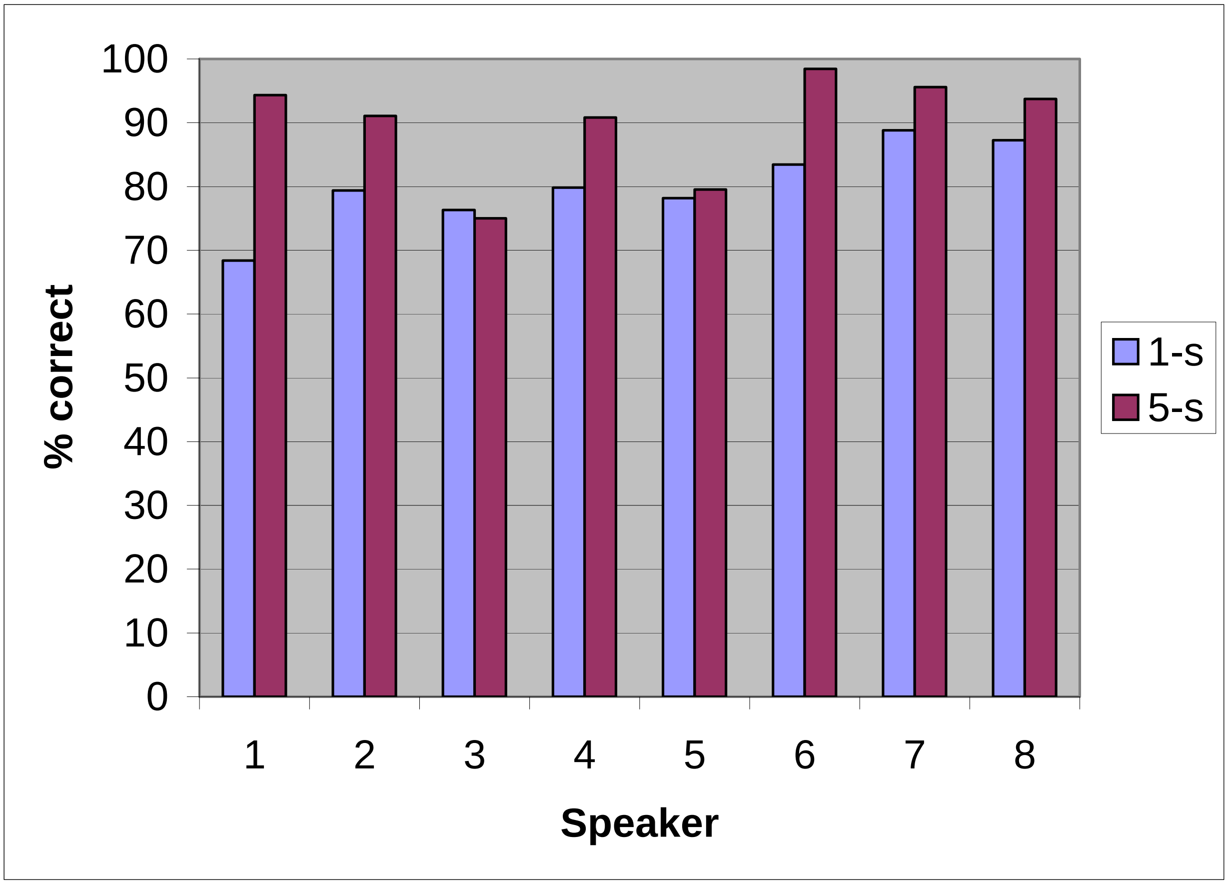
Responses from 5-s intervals were converted to responses to 1-s intervals so that the results on different interval lengths could be compared directly. To do this, 1) a 5-s FLU interval was considered to be made up of five 1-s FLU intervals, and 2) a 5-s STUT interval was considered to be made up of five 1-s STUT intervals. The second assumption operationalizes the view that too much material is designated STUT when longer intervals are used (i.e., the main topic addressed in this paper). After this response translation, comparison can be made between intervals of different length. Table 3 gives the mean percentage correct (and SD) for the naïve judges separately for each speaker and separately for both interval lengths using the experts' responses to the agreed intervals as the criterion (more extensive data from individual judges are given in Table A.1 at the end of this article).

Table 3. Mean percentage correct and standard deviation across judges for 1-s and 5-s intervals and for each speaker.

	1-s		5-s	
Speaker	Mean	SD	Mean	SD
S1	68.41	6.87	94.32	4.70
S2	79.40	2.83	91.07	5.37
S3	76.33	3.62	75.00	4.71
S4	79.83	3.93	90.83	4.95
S5	78.18	5.99	79.55	4.42
S6	83.44	5.17	98.44	4.42
S7	88.82	2.52	95.59	5.21
S8	87.25	3.37	93.75	5.18

The mean percentages for each speaker from Table 3 are presented in histogram form in Figure 1. Speakers are indicated on the abscissa, blue bars represent 1-s judgements and red bars represent 5-s judgements. It can be seem that for all but one speaker (S3), there were more 5-s intervals judged STUT than 1-s intervals judged STUT. Thus for seven out of eight of the speakers, the judgments about STUT across 1-s and 5-s intervals were in the expected direction according to prediction one.

Figure 1. Percent correct judgements of naïve judges (relative to experts' agreed responses). Results are shown for individual speakers (labeled along the abscissa) separately for 1-s and 5-s TIs.



To test the first prediction statistically, STUT intervals alone were examined to see whether more TIs were judged STUT when the longer (5-s) intervals were used than when the shorter (1-s) intervals were used. A mixed model Analysis of Variance was employed with the within-groups factor of interval length (1-s versus 5-s) and the between-groups factor of speaker (the eight speakers whose speech was judged) and the dependent variable was proportion of intervals judged stuttered). There was a significant effect of interval length ($F(1,56) = 20.4$, $p < .001$) which arose because more speech was judged STUT when 5-s intervals were used than when 1-s intervals were used. There was also a significant effect of speaker ($F(7,56) = 5.410$, $p < .001$) which indicated total stuttering rate (across 1-s and 5-s TI) differed (i.e., showed the speakers differed in severity of stuttering). There was also an interaction between interval length and speaker ($F(7,56) = 3.27$, $p < 0.01$). This arose because the effect on stuttering rate of changing interval length (stuttering rate increase going from 1-s to 5-s intervals) depended on speaker. Inspection of individual speaker data revealed that more severe stutters showed less effect than milder ones. This effect is explored in more detail in the next section.

Prediction two

The second prediction tested was that the participants who had a more severe stutter had less chance of losing intervals than milder ones. This prediction was based on the fact that the first two exclusion criteria in Table 1 would apply less to speakers with a severe stutter, as few of their intervals did not include a real stutter before and after these exclusion criteria were applied. Essentially this implies that the first two exclusion criteria were less applicable to speakers with a severe stutter than speakers with a milder stutter. This predicts that there should be a negative correlation between amount of speech lost from the expert judges and the percentage of TIs judged STUT after the exclusion criteria were applied (i.e. a one-tail prediction). The Pearson product moment correlation coefficient was -.512 which was in the correct direction but not significant (p = .10, one tail) which is not surprising given the small N. The related correlation coefficient between amount of speech lost and total length of those 5-s intervals designated STUT, correlated negatively r = -.818, p = .013 with an N of 8. This gives qualified support to the view that speakers with a less severe stutter lose more intervals due to the exclusion criteria than speakers with a more severe stutter. Thus, TI assessments using 5-s intervals affect speakers with different stuttering severity differentially.

Examination of data from individual judges and speakers

The data for the individual judges for each speaker are given in Table A.1. These data show which length intervals are judged more consistently by the naïve judges, and some additional information concerning variability between judges. Looking at interval length first, the right-most section gives the proportion of the total correct responses the naïve judges made (relative to the experts) separately for 1-s (first column of this section) and 5-s (middle column of this section) intervals and the signed difference between the two (5-s – 1-s). The majority of the latter signed differences are positive, which indicates that naïve judges were more consistent with the experts for long intervals.

Looking at variability across interval length (Table A.1), there appears to be higher numbers of false positives (i.e., calling FLU intervals STUT) in naïve judges' ratings of the 1-s than the 5-s intervals. For example, judge four assessing speaker three had 29 out of the total 75 1-s intervals rated as STUT, which included 17 false positives. So 17/29 (more than 58%) of this judge's STUT responses are wrong. For 5-s intervals for this same speaker and judge, 11 out of the total 15 intervals were rated as STUT, including three false positives. So 3/11 (27.27%) of this judges STUT responses were false positives for the 5-s intervals. 48 speaker by judge sets of data were available for whom this calculation was possible (there were no agreed fluent intervals judged for speakers two and four). For these data, the percentage of false positives was 52.6% for the 1-s intervals and 16.9% for the 5-s intervals which was highly significant by related t test ($t(47) = 10.5$, $p < .001$) Thus judgements about 1-s intervals are more prone to false positive STUT responses than 5-s intervals.

4. Discussion

The main result is that estimated stuttering rate depends on interval length with longer intervals more likely to be judged STUT. The second main finding was that the effect of interval size depended on the speaker's stuttering severity (more severe stutterers tend to have more intervals consistently judged 'stuttered' than milder ones). The experiment controlled for decison context between the expert judges that provided the criteria responses and the naïve judges by having both sets of judges assess all materials. If only expert-agreed intervals had been assessed, different range and frequency effects would have applied to the different materials and this would have affected the responses given (Parducci, 1965). The results suggest that users of TI procedures should not have free choice over interval length; otherwise the results across clinics and with different clients are not comparable. Moreover use of long intervals (5-s and over) is not recommended for detecting changes in stuttering frequency across conditions as the procedures are insensitive even to large changes in stuttering rate (this applies to Ingham et al., 1997).

The study points to major issues of reliability of the TI procedure such as difficulty in selecting a good number of samples, particularly in fluent speech, with the rejection criteria being as they are; large proportion of false positives; inability of longer intervals to measure differences in more severe stuttering. Although TI procedures can be automated and would then provide a relatively efficient method for assessing speech, these problems rule out using these procedures in clinics. TI procedures may have other uses with respect to stuttering, however. The method as applied in the current study has provided intervals which are completely fluent or contain one type of stuttering. These could be used for training and testing material for procedures that automatically count stuttering events (Howell & Huckvale, 2004). Also, the intervals can be used to establish what acoustic information is salient for detecting stutterings.

Acknowledgement. This research was supported by programme grant 072639 awarded by the Wellcome Trust.

References

Cordes, A. K. (1994). The reliability of observational data: I. Theories and methods for speech-language pathology. *Journal of Speech and Hearing Research*, **37**, 264-278.

Curlee, R. F. (1981). Observer agreement on disfluency and stuttering. *Journal of Speech and Hearing Research*, **24**, 595-600.

Howell, P., & Huckvale, M. (2004). Facilities to assist people to research into stammered speech. *Stammering Research*, **1**, 130-242.

Howell, P., Sackin, S., & Glenn, K. (1997). Development of a two-stage procedure for the automatic recognition of dysfluencies in the speech of children who stutter: I. Psychometric procedures appropriate for selection of training material for lexical dysfluency classifiers. *Journal of Speech, Language and Hearing Research*, **40**, 1073-1084.

Howell, P., Staveley, A., Sackin, S., & Rustin, L. (1998). Methods of interval selection, presence of noise and their effects on detectability of repetitions and prolongations. *Journal of the Acoustical Society of America*, **104**, 3558-3567.

Ingham, R. J., Cordes, A. K., & Gow, M. L. (1993). Time-interval measurement of stuttering: Modifying interjudge agreement. *Journal of Speech and Hearing Research*, **36**, 503-515.

Ingham, R. J., Moglia, R. A., Frank, P., Costello-Ingham, J., & Cordes, A, (1997). Experimental investigation of the effects of frequency-altered feedback on the speech of adults who stutter. *Journal of Speech, Language and Hearing Research*, **40**, 361-372.

Martin, R. R., & Haroldson, S. K. (1981). Stuttering identification: Standard definition and moment of stuttering. *Journal of Speech and Hearing Research*, **24**, 59-63.

Parducci, A. (1965). Category judgement: A range-frequency model. *Psychological Review*, **72**, 407-418.

Perkins, W.H. (2001). Stuttering: A matter of bad timing. *Science*, **294**, 786.

[1.] The interval starts and ends (both 1-s and 5-s) are imposed irrespective of where utterances start. Providing interval durations are as long as the utterance duration or longer, there will be no effect due to the tendency of stutterings to occur at the beginning of sentences,
[2.] The expert judges also indicated which of the intervals contained part-word repetitions or prolongations alone. These are intended for use in future studies and are not reported here as the procedures and treatment of results parallel those that Ingham and co-workers used in their studies.

Table A.1. The three sections (going from left to right) give results for 1-s, and for 5-s, TIs and comparisons between 5-s and 1-s TIs. Speaker (S1-S8) and naïve judge (J1-J8) are labeled at the left of each row. The three rows under the two sections headed 'Correct/total 1-s' and 'Correct/total 5-s', are, form left to right, number of responses the naïve judge got correct relative to the expert for a) fluent TI (FLU), b) stuttered TI (STUT) and c) all TI (A). The section headed 'Comparison of 5-s and 1-s' gives (going left to right), the total proportion of TI that were judged correct for 1) 1-s, and 2) 5-s intervals and 3) the signed difference between the 5-s and 1-s proportions.

	Correct/total 1-s			Correct/total 5-s			Comparison 5-s and 1-s		
	FLU	STUT	A	FLU	STUT	A	PropA 1-s	PropA 5-s	DiffA (5-s - 1-s)
S1J1	39/49	3/6	42/55	8/8	3/3	55/55	0.76	1.00	0.33
J2	33/49	4/6	37/55	8/8	2/3	50/55	0.67	0.91	0.24
J3	39/49	3/6	42/55	7/8	3/3	50/55	0.76	0.91	0.15
J4	34/49	3/6	37/55	8/8	3/3	55/55	0.67	1.00	0.33
J5	35/49	4/6	39/55	8/8	3/3	55/55	0.71	1.00	0.29
J6	34/49	3/6	37/55	8/8	2/3	50/55	0.67	0.91	0.14
J7	33/49	4/6	37/55	7/8	3/3	50/55	0.67	0.91	0.14
J8	26/49	4/6	30/55	8/8	2/3	50/55	0.55	0.91	0.36
S2J1	33/42	52/63	85/105		20/21	100/105	0.81	0.95	0.14
J2	35/42	53/63	88/105		19/21	95/105	0.84	0.90	0.06
J3	31/42	47/63	78/105		17/21	85/105	0.74	0.80	0.06
J4	33/42	50/63	83/105		20/21	100/105	0.79	0.95	0.16
J5	34/42	50/63	84/105		19/21	95/105	0.80	0.90	0.10
J6	33/42	52/63	85/105		20/21	100/105	0.81	0.95	0.14
J7	32/42	49/63	81/105		20/21	100/105	0.77	0.95	0.18
J8	33/42	50/63	83/105		18/21	90/105	0.79	0.86	0.07
S3J1	46/60	13/15	59/75	4/6	7/9	55/75	0.79	0.73	0.06
J2	50/60	12/15	62/75	5/6	5/9	50/75	0.83	0.67	0.16
J3	43/60	11/15	54/75	4/6	7/9	55/75	0.72	0.73	0.01
J4	43/60	12/15	55/75	3/6	8/9	55/75	0.73	0.73	0.00
J5	44/60	12/15	56/75	5/6	7/9	60/75	0.75	0.80	0.05
J6	46/60	13/15	59/75	5/6	6/9	55/75	0.79	0.73	0.06
J7	45/60	13/15	58/75	6/6	6/9	60/75	0.77	0.80	0.03
J8	43/60	12/15	55/75	5/6	7/9	60/75	0.73	0.80	0.07
S4J1	32/39	30/36	62/75		13/15	65/75	0.83	0.87	0.04
J2	30/39	26/36	56/75		15/15	75/75	0.75	1.00	0.25
J3	31/39	30/36	61/75		13/15	65/75	0.81	0.87	0.06
J4	33/39	30/36	63/75		13/15	65/75	0.84	0.87	0.03
J5	29/39	26/36	55/75		14/15	70/75	0.73	0.93	0.20
J6	31/39	28/36	59/75		14/15	70/75	0.79	0.93	0.14
J7	32/39	30/36	62/75		13/15	65/75	0.83	0.87	0.04
J8	31/39	30/36	61/75		14/15	70/75	0.81	0.93	0.12
S5J1	39/53	0/2	39/55	7/10	1/1	40/55	0.71	0.73	0.02
J2	47/53	2/2	49/55	9/10	1/1	50/55	0.89	0.91	0.02
J3	43/53	1/2	44/55	8/10	1/1	45/55	0.80	0.82	0.02
J4	42/53	1/2	43/55	7/10	1/1	40/55	0.78	0.73	-0.05
J5	40/53	2/2	42/55	8/10	1/1	45/55	0.76	0.82	0.06
J6	40/53	1/2	41/55	8/10	0/1	40/55	0.75	0.73	-0.02

J7	39/53	1/2	40/55	7/10	1/1	40/55	0.73	0.73	0.00
J8	44/53	2/2	46/55	9/10	1/1	50/55	0.84	0.91	0.07
S6J1	26/34	6/6	32/40	4/4	4/4	40/40	0.80	1.00	0.20
J2	30/34	6/6	36/40	4/4	4/4	40/40	0.90	1.00	0.10
J3	27/34	6/6	33/40	4/4	4/4	40/40	0.83	1.00	0.17
J4	28/34	6/6	34/40	4/4	4/4	40/40	0.85	1.00	0.15
J5	25/34	5/6	30/40	4/4	¾	35/40	0.75	0.88	0.13
J6	26/34	6/6	32/40	4/4	4/4	40/40	0.80	1.00	0.20
J7	30/34	6/6	36/40	4/4	4/4	40/40	0.90	1.00	0.10
J8	28/34	6/6	34/40	4/4	4/4	40/40	0.85	1.00	0.15
S7J1	46/53	29/32	75/85	2/3	14/14	80/85	0.88	0.94	0.06
J2	49/53	27/32	76/85	3/3	14/14	85/85	0.89	1.00	0.11
J3	48/53	29/32	77/85	2/3	13/14	75/85	0.91	0.93	0.02
J4	46/53	29/32	75/85	3/3	14/14	85/85	0.88	1.00	0.12
J5	49/53	29/32	78/85	3/3	14/14	85/85	0.92	1.00	0.08
J6	44/53	27/32	71/85	3/3	12/14	75/85	0.84	0.93	0.09
J7	46/53	29/32	75/85	2/3	14/14	80/85	0.88	0.94	0.06
J8	48/53	29/32	77/85	3/3	14/14	85/85	0.91	1.00	0.09
S8J1	28/33	14/17	42/50	0/1	9/9	45/50	0.84	0.90	0.06
J2	30/33	14/17	44/50	1/1	8/9	45/50	0.88	0.90	0.02
J3	28/33	13/17	41/50	1/1	8/9	45/50	0.82	0.90	0.08
J4	30/33	13/17	43/50	1/1	8/9	45/50	0.86	0.90	0.04
J5	29/33	16/17	45/50	0/1	9/9	45/50	0.90	0.90	0.00
J6	30/33	13/17	43/50	1/1	9/9	50/50	0.86	1.00	0.14
J7	29/33	16/17	45/50	1/1	9/9	50/50	0.90	1.00	0.10
J8	30/33	16/17	46/50	1/1	9/9	50/50	0.92	1.00	0.08

Note: As indicated in the text, there were cases where 5-s intervals were agreed to be STUT by the expert judges which contained one or more 1-s intervals that were agreed by the experts to be FLU (though at least one 1-s interval was expert-agreed STUT). This arose when the experts agreed that there was a stuttering of less than 5-s in length. This results in agreed FLU 1-s TI from the expert judges for subjects 2 and 4 where there were no agreed FLU 5-s TIs.

The impact of word-end phonology and morphology on stuttering

Chloe Marshall
Centre for Developmental Language Disorders and Cognitive Neuroscience, Department of Human Communication Science, University College London, Chandler House, 2 Wakefield Street, London WC1N 1PF
c.marshall@ucl.ac.uk

Abstract. This paper investigates whether stuttering rates in English-speaking adults and children are influenced by phonological and morphological complexity at the ends of words. The phonology of English inflection is such that morphological and phonological complexity are confounded, and previous research has indicated that phonological complexity influences stuttering. Section 1 of this paper considers how to disentangle phonological and morphological complexity so that the impact of each on stuttering can be tested. Section 2 presents an analysis of some adult corpus data, and shows that phonological and morphological complexity at the word end do not influence stuttering rates for English-speaking adults, at least in spontaneous speech. Section 3 presents results from a non-word repetition task and a past tense elicitation task which reveal that while word-end phonological and morphological complexity do not affect stuttering rates in most of the adults and children tested, a small proportion of adults and children do stutter over morphologically complex words in an elicitation task. Taken as a whole, these results suggest that morphology has an impact on stuttering for some individuals in certain circumstances. **Keywords:** Developmental stuttering, word-end phonology, word end morphology.

1.1 Introduction

Several recent models of stuttering hypothesise that the linguistic characteristics of the word being attempted can trigger stuttering (e.g. Au-Yeung & Howell, 1998; Packman, Onslow, Richard & van Doorn, 1996). While the role of phonological factors in stuttering has received some attention in the literature, the impact of word-end phonology has not been studied independently from other aspects of phonology. Nor has the role of inflectional morphology (plural marking on nouns; tense, agreement and aspectual marking on verbs) in stuttering been investigated. Word-end phonology and inflectional morphology are considered together in this paper for the following reason: In English, inflectional morphology occurs at the ends of words, and changes the phonology of the word end. For example, the plural form of '*cat*', *cats*, ends in a cluster, whereas the singular form, *cat*, does not. Similarly the verb '*lie*' ends in a vowel in the forms *I lie* and *you lie*, but in a consonant in the past tense forms *I lied* and *you lied*. As a consequence of this close relationship between phonology and morphology, the impact of morphology on the production of words is best studied alongside that of phonology.

The research questions addressed in this paper are as follows: (1) does phonological complexity at the ends of words influence stuttering rates, (2) does inflectional morphology influence stuttering rates, and (3) if morphology does influence stuttering, is this effect independent of phonological complexity? The paper attempts to respond to Demuth's plea that the role of phonology and morphology in stuttering be investigated (Demuth, 2004).

1.2 The impact of phonology on stutttering

Researchers use differing measures of what constitutes phonological or phonetic complexity, and therefore they differ in what phonological and phonetic characteristics they consider are likely to cause stuttering. As a rule these measures are based on what typically developing children find hard to acquire. For example, Throneburg, Yairi and Paden (1994) investigated the impact of late developing sounds, consonant clusters and multisyllabicity. A more comprehensive metric is the Index of Phonetic Complexity (IPC; Jakielski, 1998; Weiss & Jakielski, 2001), whereby words are scored according to how many 'difficult' structures they contain, with the most difficult words being those that get a high score. The IPC is outlined in Table 1 below.

Table 1. Index of Phonetic Complexity (Jakielski, 1998)

Factor	No score	One point each
1. consonant by place	Labials, coronals, glottals	dorsals
2. consonant by manner	stops, nasals, glides	fricatives, affricates, liquids
3. singleton consonants by place	Reduplicated	variegated
4. vowel by class	monophthongs, diphthongs	rhotics
5. word shape	ends with a vowel	ends with a consonant
6. word length (syllables)	monosyllables, disyllables	three or more syllables
7. contiguous consonants	no clusters	consonant clusters
8. cluster by place	Homorganic	heterorganic

Results from these studies have been mixed. Throneburg et al. (1994) found that neither late-developing sounds, consonant clusters nor multisyllabicity had any effect on stuttering rates in children aged between two and a half and five years. Howell and Au-Yeung (1995) confirmed this finding for a wider range of ages, up to twelve years old.

The IPC has been used in various investigations (Dworzynski & Howell, 2004; Howell, Au-Yeung, Yaruss & Eldridge, submitted; Weiss & Jakielski, 2001). Weiss and Jakielski (2001) studied children between the ages of six and eleven and a half. They found a trend for prosodic complexity to influence stuttering rates more in younger children than in older children, although this difference was not significant. In contrast, Howell et al. (submitted) found that phonetic complexity only influenced stuttering rates in children older than eleven and in adults. Studies using the IPC have revealed different patterns cross-linguistically. For example, Dworzynski and Howell (2004) found that for German people who stutter (PWS), words ending in consonants are more likely to be stuttered than words ending in a vowel, and that this effect was present for both adults and children over the age of six. This effect was not found for English speakers (Howell et al., submitted), a finding that Dworzynski and Howell (2004) put down to the fact that words ending in consonants are more frequent in English than in German.

One of the disadvantages of the IPC is that, apart from the factor 'word shape' (i.e. whether the word ends in a consonant or a vowel), it does not specify where in the word phonetic complexity occurs. It is well-established that word-initial sounds trigger disfluencies (e.g. Conture, 1990; Howell, Au-Yeung & Sackin, 2000; Natke, Sandreiser, van Ark, Pietrowski & Kalveram, 2004; Wingate, 1982, 1988), but the aim of this paper is to investigate the impact of the word end. Dworzynski and Howell (2004) suggest that word-final factors might indeed play a role in the planning and retrieval time of words, given that word shape affects German PWS. Otherwise, there is as yet little evidence that word-end phonology influences stuttering.

1.3 The relationship between phonology and inflectional morphology

In this paper I consider the impact not only of phonological complexity at the word end, but also that of inflectional morphology. English morphology is sparse, with a mere two forms for nouns (*cat*, *cats*) and a maximum of five for verbs (*sew*, *sewed*, *sews*, *sewing*, *sewn*). I am concerned only with suffixes that add a consonant – past tense *τ/δ* (e.g. *hoped*, *sewed*), plural *σ/ζ* (e.g. *cats*, *dogs*) and third person singular *σ/ζ* (e.g. *hopes*, *sews*). By adding a consonant, these suffixes increase phonological complexity at the word end, and may create clusters (compared *sewed* with *hoped*). I leave aside the syllabic suffixes -*εδ* (e.g. *wanted*, *needed*), -*εζ* (e.g. *horses*, *reaches*) and -*ing*.

Any theory of phonological and morphological complexity should mirror what we know happens developmentally. Children's first words end in a vowel (e.g. *mama*) and only later, at about the age of one and half or two do they produce words ending in a consonant, e.g. *man*. Only later still will they produce words ending in a consonant cluster. Therefore word-final clusters are more complex in some way than word-final singleton consonants. In terms of children's early word productions, their first words are generally uninflected, and only later are inflections produced. *-Ing* and plural *–s* are among the first inflections used, while third person singular *–s* and past tense *–ed* take longer to become reliably established (see Bernhardt & Stemberger, 1998, for an overview of phonological development).

Howell et al. (submitted) investigated the effect that a word-end consonant has on stuttering rates in adults, and found that words ending in a consonant were actually stuttered less frequently than those ending in a vowel. The first indications would therefore appear to be that word-end morphology that adds a consonant might not actually be relevant to stuttering, at least in adults. However, the possibility that morphology does affect stuttering cannot be ruled out on the basis of Howell et al.'s (submitted) data. To do so, a systematic analysis of phonological and morphological factors at the word end is needed.

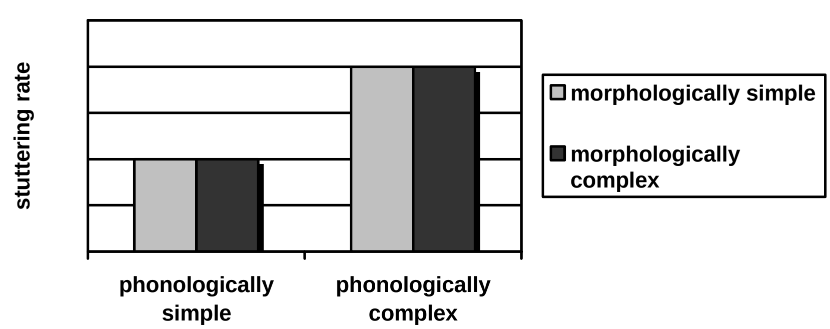
In the analyses presented in this paper, word forms that end in a singleton consonant are termed 'phonologically simple'[1] and those that end in a cluster 'phonologically complex'. Uninflected words are termed 'morphologically simple' and inflected words 'morphologically complex'. This classification enables four types of words to be contrasted:

1. phonologically simple, morphologically simple (e.g. *bad*, *mouse*)
2. phonologically simple, morphologically complex (e.g. *died*, *days*)
3. phonologically complex, morphologically simple (e.g. *month*, *think*)
4. phonologically complex, morphologically complex (e.g. *looked*, *wants*)

Such a classification enables the following questions to be raaised: (1) does phonological complexity at the ends of words influence stuttering rates independent of morphology, and (2) does inflectional morphology influence stuttering rates independent of phonology?

The pattern of results that might plausibly be expected to arise are now discussed. If phonological complexity at the ends of words influences stuttering rates independently of morphology, the following pattern of results would be predicted – higher stuttering rates on 3+4 (i.e. phonologically complex words) than on 1+2 (i.e. phonologically simple words), but no difference between 1 and 2 (i.e. between morphologically simple and complex words that are phonologically simple) and no difference between 3 and 4 (i.e. between morphologically simple and complex words that are phonologically complex) (see Figure 1a). If morphological complexity influences stuttering rates then higher stuttering rates would be seen on 2+4 (i.e. morphologically complex words) than on 1+3 (i.e. morphologically simple words). If morphological complexity influences stuttering independently of phonology, then stuttering rates would be expected to be equivalent for 2 (i.e. phonologically simple) and 4 (i.e. phonologically complex) (see Figure 1b). A third possibility is that there may be an interaction between phonological and morphological complexity, with stuttering rates highest for those words that are both phonologically and morphologically complex (Figure 1c).

Figure 1a. Phonological complexity influences stuttering independent of phonology



[1] Of course, even more phonologically simple would be a word that ends in a vowel rather than a consonant, but there are no inflected words of this shape with which to compare it.

Figure 1b. Morphological complexity influences stuttering independent of phonology

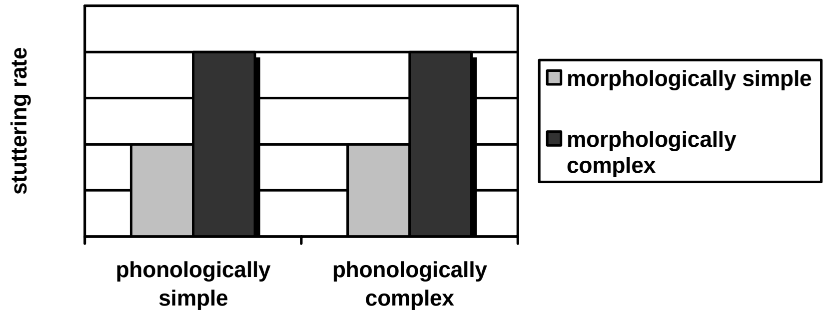
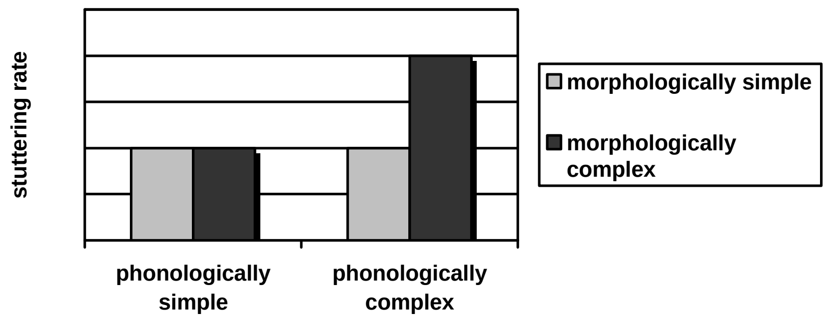


Figure 1c. Interaction between phonological and morphological complexity



In order to investigate which of these patterns of performance is found amongst PWS, two types of data are considered. In Study 1, a corpus of teenage and adult speech is analysed, looking at the patterns of relative performance on phonologically and morphologically complex forms (as outlined above). In Studies 2 and 3, some experimental data from children and young adults, from a non-word repetition task and from a task designed to elicit inflected verbs are analysed. The non-word repetition task allows phonologically simple and complex non-words to be compared in a context that is by its very nature morphologically simple. The elicitation task allows comparison between phonologically simple and complex words in the presence of morphological complexity.

2.1 Study 1.

This study is an analysis of words selected from the University College London Archive of Stuttered Speech (UCLASS) data base (Howell & Huckvale, 2004). The material was produced spontaneously in response to prompt questions by an interviewer and conforms to “casual” speech in terms of Labov’s (1978) stylistic continuum in sociolinguistics. The speech was recorded from 16 male participants, diagnosed as showing signs of stuttering by a speech pathologist in a clinic in the United Kingdom. They were then independently diagnosed by a second speech pathologist as a person who stutters. They ranged in age from 16 to 47 years, with a mean age of 24.93 years (sd = 9). Number of words in the sample, the percentage of these words that were stuttered, and their age at the time of recording are given for each participant in Table 2 below. The samples of speech were heterogeneous so as to reflect a wide variety of stuttering profiles. The material was transcribed according to the conventions outlined in Howell and Huckvale (2004). The transcriptions indicate where stuttering was located (see Kadi-Hanifi & Howell, 1992, for an indication what events were considered to be stutterings and estimates of reliability for this type of material).

Table 2. Details of speakers, total number of words in the sample, number of words stuttered, percentage stuttering rate in the sample and age in years

Participant number	Number of words	Number of words stuttered	% stuttered	Age
1	1420	37	2.61	47
2	1720	46	2.70	25
3	1016	31	3.10	38
4	1255	61	4.90	26
5	1390	76	5.47	28
6	188	11	5.90	24
7	219	14	6.40	18
8	221	15	6.80	16
9	338	23	6.80	17
10	1315	125	9.51	20
11	272	28	10.30	16
12	763	85	11.10	25
13	1593	179	11.24	21
14	378	43	11.40	16
15	943	154	16.33	35
16	1082	220	20.33	40

In this analysis stuttering rates are compared across four different types of words, whose characteristics (and some examples) are set out in Table 3 below. Phonologically, words differ only in whether they contain a word-final cluster. The words selected have either the shape Consonant-Vowel-Consonant (CVC) or Consonant-Vowel-Consonant-Consonant (CVCC), i.e. none contain onset clusters, and none consist of more than one syllable.

Table 3. Characteristics of the words being analysed and some examples

		Morphology	
		simple	complex
phonology	simple	*bad,* *house*	*died,* *days*
	complex	*month,* *think*	*looked,* *wants*

A Perl program (Brown, 2001) identified words ending in the letters s, z ,d and t. These were then hand-coded as being morphemic or otherwise. Irregular verbs were then removed from the data set, again by hand. These included irregular past tenses (e.g. *hit*, *had*, *got*, *was*), irregular third person singular *has*, and past participles (e.g. *done*). They were removed because irregulars are not strictly speaking decomposable into verb stem + suffix in the way that regular verbs are (e.g. *lies* = *lie* + *s* but *has* ≠ *ha* + *s*; therefore *lies* is clearly morphologically complex, but *has* is not).

Scoring

Each word was classified as stuttered (1) or not stuttered (0), based on such events marked in the original transcriptions (see Howell & Huckvale, 2004 for details).

2.2 Results

Table 4 shows the total number of words of each type that were extracted from the database, and then in brackets the number that were stuttered.

Table 4. Total number of words in each type (number of stuttered words in brackets); M = mean number of words stuttered, expressed as a percentage

		morphology	
		simple (uninflected)	complex (inflected)
Phonology	simple (CVC)	509 (48) M = 9.43%	23 (2) M = 8.70%
	complex (CVCC)	94 (7) M = 7.45%	190 (16) M = 8.42%

Overall, 8.95% of words were stuttered. Differences in the percentage of stuttered items in each cell were very small, and chi-square analysis shows them to be statistically insignificant (for phonologically simple words, Cramer's V = 0.005, p = 0.906; for phonologically complex words, Cramer's V = 0,017, p = 0.777; for morphologically simple words, Cramer's V = 0.025, p = 0.539; for morphologically complex words, Cramer's V = 0.003, p = 0.964).

Ideally it would be desirable to analyse individual performance, in order to determine whether any speakers were significantly affected by phonological and/ or morphological complexity. However, this analysis was not possible because the speakers produced greatly different numbers of words in the different categories of interest, and some speakers produced no tokens within a particular category, either fluent or stuttered.

From these results it can be concluded that for English-speaking adults and adolescents the presence of a cluster at the end of a word does not influence stuttering rates in natural speech. Nor is there an effect of inflectional morphology on stuttering. However, it is stressed that the data are not amenable to individual analysis because often speakers did not produce any tokens within a particular category. The possibility remains that for certain speakers either or both types of complexity might have an effect. Experimental studies provide an opportunity to elicit a larger number of tokens of the particular word shapes that are of interest.

3.1. Studies 2 and 3: Experimental data from children and young adults.

In this section two experimental studies are reported. Study 2 uses a repetition task, which is designed to elicit nonsense words with or without clusters at the word end. Study 3 uses an elicitation task, and is designed to elicit third person singular (e.g. *weighs*, *wraps*) and past tense forms (e.g. *weighed*, *wrapped*) with or without word-final clusters. These tasks permit an investigation of the impact of phonological complexity on stuttering rates, first within morphologically simple, and then within morphologically complex, words.

As was seen in Study 1, samples of spontaneous speech might not contain enough tokens of the sort that are of interest, so that even if group analyses are possible, individual analyses are not. Elicitation tasks have been used previously for investigating syntactic effects on stuttering. For example, Silverman and Bernstein Ratner (1997) asked children to repeat sentences with differing types of complex syntactic structures, such as WH questions and centre embedded sentences. A standard technique used in the atypical phonological development literature involves children playing with toys that have particular names so that the experimenter can elicit the phonological forms of interest (e.g. Chiat, 1989). This technique offers a halfway house between free spontaneous speech and formal elicitation methods, but as far as is known, it has not been used for studying the phonological development of children who stutter (although it has been used to ascertain whether such children have awareness of their speech problem, Yairi & Ambrose, 2004).

Experimental studies allow balanced numbers of the types of words that are of interest to be gathered, controlled for factors such as lexical frequency, in a short period of time. A disadvantage is that if very different results are obtained as compared to analyses of spontaneous speech, then these would need to be interpreted. For example, if higher rates of stuttering were discovered for certain classes of words in experimental studies compared to spontaneous speech, is that because the experimental task is forcing speakers to make errors that they wouldn't normally make, or is it because in spontaneous speech speakers can avoid using words that they are aware they might stutter over? And yet, even if errors are forced in an experimental situation, that can still give us valuable information about the cognitive processes underlying stuttering. The merits of analyzing spontaneous material versus material obtained experimentally in young children have been discussed at length by Savage and Lieven (2004), who conclude that both are essential research tools.

3.2.1 Study 2 – non-word repetition

In non-word repetition tasks the participant hears nonsense words and repeats them. These tasks have been widely used with children with specific language impairment (SLI) and dyslexia. Children with SLI and dyslexia typically make repetition errors as the non-words get longer (Bishop, North & Donlan, 1996; Gathercole & Baddeley, 1990; Martin & Schwartz, 2003). Controversy still exists over the precise locus of the deficit that gives rise to non-word repetition difficulties. The evidence was initially interpreted as revealing that children with SLI have verbal short-term memory limitations, but this interpretation has been challenged from many quarters.

Marshall, Harris and van der Lely (2003) contend that non-word repetition tasks can be used to probe the status of phonological skills in children. Manipulating the phonological structure of non-words allows investigation of which structures are easy to repeat (and hence which are easily represented/ processed) and which are hard to repeat (less easily represented/ processed). Marshall et al. (2003) and Gallon, Harris and van der Lely (submitted) used the Test of Phonological Structure (TOPhS, van der Lely & Harris, 1999) to reveal that SLI children have difficulty repeating non-words that contain clusters and unfooted syllables (a real word example of an unfooted syllable would be the initial unstressed syllable in *gorilla*). Note that the children in van der Lely and colleagues' studies did not have verbal dyspraxia or disfluencies – their articulation was clear and fluent for known words.

The aim of the non-word repetition task included here is to investigate whether phonological complexity at the end of nonwords, as indexed by the presence of a consonant cluster, influences stuttering rates. Note that other studies of non-word repetition (e.g. Hakim & Bernstein Ratner, 2004) have investigated both whether PWS make more repetition errors than people who do not stutter, and whether stuttering rates increase as the length of the non-word increases. Hence they have investigated both the non-word repetition accuracy of PWS and whether the properties of non-words trigger stuttering. In this analysis, interest is solely in the latter question.

3.2.2 Method

Participants

19 speakers participated, and their details are presented in Table 5. Their average age is 14;0. For the purposes of analysis they were divided into 2 groups: the participants in Group 1 are aged 14 years and younger, and those in Group 2 are 15 and older. The motivation for dividing the group at 14 years of age was that around this age, stuttering changes form (the type of stuttering that occurs and the words on which the stuttering is located differ from before to after teenage). Thus the two age groups might operate differently with respect to stuttering.

Table 5. Details of speakers who participated in the experimental tasks

Participant number	Age	Group	Gender
9	8;0	1	M
10	9;10	1	M
11	9;8	1	M
12	10.5	1	M
14	10;5	1	F
15	10;10	1	M
13	11;4	1	M
3	12;10	1	M
1	13;5	1	M
16	14;2	1	F
19	14;4	1	F
4	14;5	1	M
18	15;5	2	M
6	15;8	2	M
17	17;3	2	F
2	17;4	2	M
5	17;7	2	F
8	19;5	2	M
7	21;10	2	M

Items

There were two experimental conditions and one filler condition, with items adapted from the TOPhS (van der Lely & Harris, 1999). For Experimental Condition 1, one-syllable items (N=8) were chosen that ended in a singleton consonant. Half of these items have a simple onset (e.g. *κEτ*) and half have a complex onset (e.g. *κλEτ*). For Condition 2, one-syllable items (N=8) were chosen that end in a two-consonant cluster. Again, half have a simple onset (e.g. *κEστ*) and half have a complex onset (e.g. *κλEστ*). The TOPhS was designed to allow investigation into the impact of word-final clusters on the repetition of non-words that are otherwise segmentally identical. This match for segmental content is rarely possible when using real word stimuli. A further point is that, by the very nature of the task, these stimuli are morphologically simple. For the Filler Condition (N = 12), three and four syllable (i.e. multisyllabic) items were chosen, none of which contains a cluster (e.g. *δE̲π≅ρι*, *σ≅πI̲φι*; stressed vowel underlined). The filler condition was chosen to contrast with the monosyllabic experimental items, with the aim of adding variety and maintaining participants' attention to the task.

Procedure

The experimenter tells the participant '*In this game I'm going to say some funny, made-up words which I would like you to repeat after me. We'll start with some practice ones so you can see what you have to do. Can you say '**ζΙκ**'?'* The experimenter gives the participant time to repeat the non-word, and repeats it if necessary. The experimenter carries on down the list of practice items, just saying the non-word with no introduction.

Before the experimenter starts the experimental items, he or she says '*Those were the practice words. You did really well with those. Now we're going to start for real.*' The experimenter says each experimental item with no introduction, and this time does not repeat any of them if the participant is unsure about them – he or she just moves on to the next item.

Scoring

Answers were scored for the presence (1) or absence (0) of stuttering. Note that repetition accuracy is not of concern in this analysis.

Predictions

If phonological complexity at the word end influences stuttering, we predict higher stuttering rates for non-words ending in a cluster.

3.2.3 Results

The results are shown in Table 6 below.

Table 6. Total number of words in each type (number of stuttered words in brackets); M = mean number of words stuttered, expressed as a percentage; SD_i = standard deviation by items; SD_p = standard deviation by participant

	Experimental conditions		**Filler condition**
Participant group	no word-final cluster (e.g. *κΕτ*, *κλΕτ*)	word-final cluster (e.g. *κΕστ*, *κλΕστ*)	multisyllabic (e.g. *σ≅πΙφι*)
1 (14 years and younger)	96 (1) M = 1.04% SD_i = 4.42 SD_p = 2.41	96 (2) M = 2.08% SD_i = 3.24 SD_p = 3.24	144 (21) M = 14.58% SD_i = 10.74 SD_p = 20.41
2 (15 years and older)	56 (1) M = 1.79% SD_i = 4.42 SD_p = 5.40	56 (1) M = 1.79% SD_i = 4.42 SD_p = 5.40	84 (11) M = 13.10% SD_i = 2.41 SD_p = 59.39

The data in Table 6 indicate that very few of the experimental items are stuttered, and that these are fairly evenly distributed between those that end in a cluster and those that do not. Obviously, with such low numbers of stuttered items, it is not possible to test for a significant difference between non-words with and without final clusters. However, multisyllabic filler items were stuttered much more frequently than the monosyllabic experimental items, and a chi-square analysis shows that across participant group this difference is significant (Cramer's V = 0.241, $p < 0.001$). Caution is expressed on interpreting these results, however, because they hide a range of individual patterns, as indicated by the very large figures for the standard deviations by participant (SD_p). Only one participant from Group 2 (participant 2) stuttered on multisyllabic words, whereas five from Group 1 (participants 3, 10, 13, 16 and 19) stuttered at least once on these items. These details are shown in Table 7.

Table 7. The number of multisyllabic items stuttered by individual speakers

Participant number	Group	Number multisyllabic items stuttered
9	1	0
10	1	4
11	1	0
12	1	0
14	1	0
15	1	0
13	1	3
3	1	2
1	1	0
16	1	5
19	1	7
4	1	0
18	2	0
6	2	0
17	2	0
2	2	11
5	2	0
8	2	0
7	2	0

As for which multisyllabic items were stuttered most, these data are shown in Table 8. As the standard deviations in Table 6 indicate, there is little variability across items.

Table 8. Multisyllabic items and how often they were stuttered

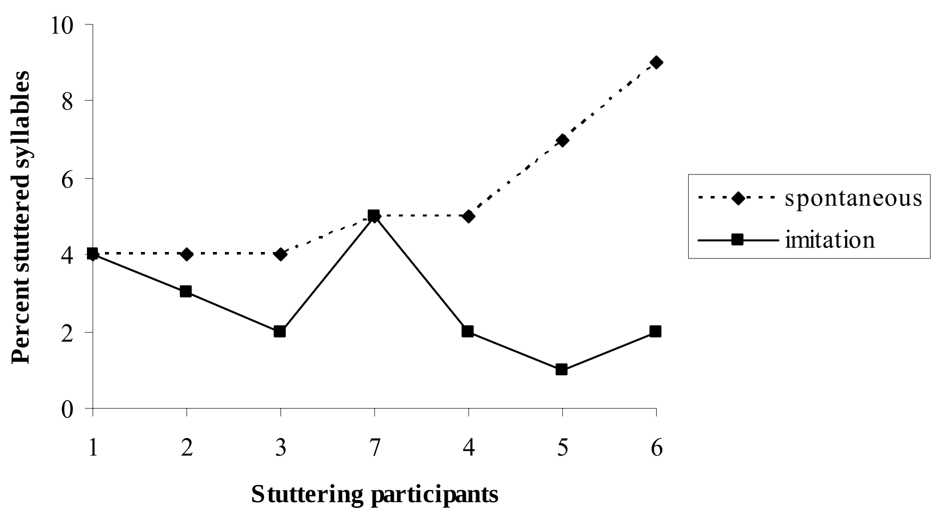
	Number of times item was stuttered	
Item	Group 1	Group 2
$\delta\underline{E}\pi\cong\rho\iota$	2	1
$\phi\underline{I}\pi\cong\lambda\cong$	0	1
$\kappa\underline{E}\tau\cong\lambda\cong$	3	1
$\pi\underline{I}\phi\cong\tau\cong$	2	1
$\beta\cong\delta\underline{E}\pi\cong$	2	1
$\delta\cong\phi\underline{I}\pi\lambda$	2	1
$\phi\cong\kappa\underline{E}\tau\varepsilon$	1	1
$\sigma\cong\pi\underline{I}\phi\iota$	0	1

β≅δ<u>E</u>π≅ρι	4	1
δ≅φ<u>I</u>π≅λ≅	0	1
φ≅κ<u>E</u>τ≅λ≅	3	0
σ≅π<u>I</u>φ≅τ≅	2	1

3.2.4 Discussion

The fact that a non-word repetition task is capable of inducing stuttering is shown by the high stuttering rates for multisyllabic items. Yet rates of stuttering are lower for those same phonological forms that were produced in spontaneous speech (see Table 3)[2]. Is there something about direct repetition of a model's speech that reduces stuttering rates? Howell and Dworzynski (in press) have raised a related question with respect to meaningful material: they examined the results from Silverman and Bernstein Ratner's (1997) study and showed that those speakers who had a higher rate of stuttering in spontaneous speech tended to have lower stuttering rates in an imitation task (see Figure 2, reproduced with their permission). Howell and Dworzynski speculate that this may arise because the participant has an opportunity of previewing and preplanning material in an imitation task. Thus, it may not only be that direct repetition reduces stuttering rate, but also, the reduction is greater in speakers who have more severe problems as measured from their spontaneous speech.

Figure 2. Percent syllables stuttered in spontaneous speech and imitation tasks for individual participants who stutter. Redrawn from Silverman and Bernstein Ratner (1997) with participants ordered in increasing percentage of syllables stuttered in spontaneous speech. Participants have the same numbers as in Silverman and Bernstein Ratner (1997).



Given the high rate of stuttering in this experiment on multisyllabic words, it would be interesting to compare this rate with the rate of stuttering for multisyllabic words in spontaneous speech samples, but that particular issue is beyond the scope of this paper.

3.3.1 Study 3

Elicitation tasks have been widely used to probe inflectional abilities, particularly in children with SLI who are known to have difficulties with inflectional morphology (Berko, 1958; Leonard, Eyer, Bedore & Grela, 1997; Rice, Wexler & Cleave, 1995; van der Lely & Ullman,

[2] We are not quite comparing like for like – half the non-words in the repetition task contain onset clusters, whereas those in spontaneous speech sample did not

2001). Many of these tasks use real verbs, but some use nonsense verbs as a way of investigating whether the child's knowledge of morphology is truly productive (as opposed to the child just supplying a memorised inflected form).

As the non-word repetition task (Study 2) gave such low stuttering rates compared to real words (as measured from spontaneous speech, Study 1), real verbs were used for this task. The aim is to determine whether phonological complexity affects stuttering rates for inflected words by comparing inflected words that end in a cluster (e.g. *wrapped*) with those that end in just a single consonant (e.g. *weighed*)

3.3.2 Method

Participants

The same speakers participated as in Study 2. See Table 5 for their details.

Items

Two experimental conditions (no fillers) were selected. Condition 1 consists of 5 verbs whose 3rd person singular and past tense forms both end in a single consonant (e.g. *weighs*, *weighed*). Condition 2 consists of five verbs whose inflected forms end in a two-consonant cluster (e.g. *wraps*, *wrapped*). The aim is to compare stuttering rates on the phonologically simple (no cluster) and phonologically complex (cluster) forms. None of the verbs has an onset cluster. Therefore the effect of complexity is investigated only at the verb end. Hubbard and Prins (1994) report frequency effects on stuttering rates, so we ensured that both conditions consisted of high frequency verbs that were matched for frequency[3].

Procedure

The items are presented verbally, and the response is requested verbally (i.e. the participant is not required to read or write anything). The experimenter says to the participant *'In this game we're going to talk about things that I like to do, and that my friend also likes to do. You're going to help me to say those things. For example, if I say 'Everyday I go to school', this is something my friend does too. So you can tell me 'Your friend also goes to school'. Can you say that? And you can also tell me what we both did yesterday: 'Yesterday you both went to school'. Can you say that?*

'Let's do another one to practice. 'Everyday I eat ice cream.' So you can say 'Your friend also (leave blank for the participant to fill in). *And you can also say 'Yesterday you both* (leave blank for the participant to fill in). *That's it. Let's do one more practise one before we start for real.'*

Most participants get the hang of this procedure very quickly. The experimenter gives the participant as much prompting as they need with the initial part of the sentence, but always lets them fill in the verb. If the participant can't remember the verb, then the experimenter repeats the first sentence (Everyday I …). Verbs are based on those used in Marshall (2004).

Scoring

Responses were scored for the presence (1) or absence (0) of stuttering. Accuracy of inflection itself is not of concern.

Predictions

If phonological complexity is a factor influencing stuttering on morphologically complex words, then more items ending in a consonant cluster would be predicted to be stuttered.

3.3.3 Results

The results are set out in Table 9 below.

---

[3] Natural log of lemma frequency, according to the English frequency count in the CELEX database (Baayen, Piepenbrock & van Rijn, 1993) was 4.5 and above.

Table 9. Total number of words in each type (number of stuttered words in brackets); M = mean number of words stuttered, expressed as a percentage; SD_i = standard deviation by items; SD_p = standard deviation by participant

	Experimental condition	
Participant group	No cluster (e.g. *weighs*, *weighed*)	Cluster (e.g. *wraps*, *wrapped*)
1 (14 years and under)	120 (14) M = 11.67% SD_i = 5.16 SD_p = 22.90	120 (14) M = 11.67% SD_i = 6.99 SD_p = 20.65
2 (15 years and older)	70 (3) M = 4.29% SD_i = 4.83 SD_p = 7.87	70 (4) M = 5.71% SD_i = 5.16 SD_p = 15.12

An examination of the data in Table 9 indicates that the presence of a cluster at the inflected verb end has no influence on stuttering rates. The standard deviations by items (SD_i) in each cell are low, reflecting that there is little variability across verbs. The only item for which none of the 19 speakers stuttered was *hums*. The items that were stuttered most were *wrapped*, *hummed* and *sews*, each stuttered by 3 of the 19 speakers.

Variability across speakers, as revealed by SD_p, is much greater, and the scores for each speaker are presented in Table 10. Only three of the twelve participants in Group 1 stutter on more than one item (participants 3, 10 and 16), and eight stutter on none of the items. One participant in Group 2 (participant 2) is responsible for all the items stuttered by his group, and he is the only one who stutters more on items containing a final cluster. Interestingly, he was the participant who stuttered on the multisyllabic non-words in the repetition task in Study 2. Although it appears that participants in the younger group stutter more on the items in the elicitation task than the older group, this difference was not tested for significance because of the large individual differences in performance in both groups.

Table 10. Number of morphologically complex items stuttered by individual speakers

Participant number	Group	Number items stuttered – no final cluster	Number of items stuttered – final cluster
9	1	0	0
10	1	7	6
11	1	0	0
12	1	0	0
14	1	0	0
15	1	0	0
13	1	0	0
3	1	4	4
1	1	0	0
16	1	3	3
19	1	0	0
4	1	0	0
18	2	0	0

6	2	1	0
17	2	0	0
2	2	2	4
5	2	0	0
8	2	0	0
7	2	0	0

3.3.4 Discussion

A few individuals stutter on morphologically complex forms, but they are in a minority. For those individuals, with the possible exception of Participant 2, there was no effect of phonological complexity upon stuttering: the presence of a cluster at the verb end did not influence stuttering rates.

4. Conclusions

The analysis of spontaneous speech (Study 1) revealed no differences in stuttering rates within words that are either phonologically complex (as indexed by the presence of a word-final cluster), morphologically complex (as indexed by the presence of inflection) or both. In a non-word repetition task (Study 2), where items are by their very nature morphologically simple, the presence of a word-end cluster has no influence on stuttering rates. In a task designed to elicit inflected words (Study 3), phonological complexity has no effect either, with the possible exception of one of the speakers in the older group. However, a small proportion of individuals do stutter on inflected forms, indicating that for certain PWS morphological complexity can affect their stuttering.

These preliminary results reveal that phonological complexity at the word end has no effect on stuttering amongst speakers of English, while morphological complexity affects only a minority of PWS. However, this is by no means the full story. For example, these studies did not consider inflections that add a syllable, e.g. *εδ*, *εζ* and *–ing*. The higher rates of stuttering on multisyllabic words in the non-word repetition test suggest that syllabic inflections might cause fluency difficulties, particularly for verb stems that are themselves more than one syllable long, e.g. *oranges*, *recorded*, *balancing*. Also beyond the scope of this paper, but perhaps worth investigating in future work, is the impact of derivational morphology, as this adds one or more syllables to either the beginning or the end of words (e.g. *heavi**er***, *muddi**est***, ***un**real*, ***sub**standard*, ***over**cater*). Certainly, on the non-word repetition task presented in Study 2, stuttering rates were high in the repetition of multisyllabic non-words with similar metrical shapes to derived words, e.g. *σ≅πIφι*, *πIφ≅τ≅*.

It would also be worth investigating the effects of cumulative complexity. It is possible that word final clusters only have a measurable effect on stuttering rates when they occur in words that contain other complex phonological structures, such as unfooted syllables and onset clusters. In support of this hypothesis, research using the IPC metric suggests that the effects of phonological complexity are indeed cumulative (e.g. Howell et al., submitted).

Nor can we conclude from the results presented here that effects of word-end phonology and of inflectional morphology will not be among the factors implicated at the onset of stuttering, which is most common at the age of 3. Tasks requiring children to repeat non-words and produce morphologically complex words have been carried out successfully with children this young (e.g. Roy & Chiat, 2004, non-word repetition; Rice & Wexler, 2001, inflection), and experimental material similar to that presented in Studies 2 and 3 of this paper would be suitable for this purpose.

Finally, one of the main aims of this paper has been to instruct the experimenter on how to tease apart phonology and morphology, so that the effects of each can be studied independently. It is my hope that it will stimulate further work in this field, improving our understanding of how word-end factors affect stuttering both at its onset and in persistent stutterers.

Acknowledgements

I would like to thank Pete Howell, for encouraging me to think about the issues raised in this paper and for his support during every stage of the writing process. Huge thanks are also due to Jenny Hayes, for writing the computer program that selected words from the corpus of spontaneous speech, and for carrying out experimental tasks. I am grateful to Heather van der Lely and John Harris for letting me use items from the Test of Phonological Structure. The bulk of this work was carried out while I was in receipt of an ESRC studentship, which is gratefully acknowledged.

Appendix 1. Non-words used in the repetition task.

Practice items

ζΙκ, β≅νΕλι, φΙν≅ρι, ωΕφ

Experimental items – no word-final cluster

φρΙπ, κΕτ , πΙφ, πρΙφ, δΕπ, δρΕπ, φΙπ, κλΕτ

Experimental items –word-final cluster

δΕμπ, κλΕστ, κΕστ, φρΙμπ, φΙμπ, δρΕμπ, πΙλφ, πρΙλφ

Filler items

δΕπ≅ρι, σ≅πΙφ≅τ≅, κΕτ≅λ≅, δ≅φΙπλ, πΙφ≅τ≅, β≅δΕπ≅ρι, β≅δΕπ≅, δ≅φΙπ≅λ≅, σ≅πΙφι, φ≅κΕτ≅ λ≅, φ≅κΕτ≅, φΙπ≅λ≅

Appendix 2. Stimuli for the third person singular and past tense elicitation task

Practice items

Everyday I go to school. Your friend also (goes) to school. Yesterday you both (went) to school.
Everyday I eat ice cream. Your friend also (eats) icecream. Yesterday you both (ate) ice cream.
Everyday I swim in the sea. Your friend also (swims) in the sea. Yesterday you both (swam) in the sea.

Experimental items – no word-final cluster

Everyday I pour a drink. Your friend also (pours) a drink. Yesterday you both (poured) a drink.
Everyday I pay a bill. Your friend also (pays) a bill. Yesterday you both (paid) a bill.
Everyday I sew a shirt. Your friend also (sews) a shirt. Yesterday you both (sewed) a shirt.
Everyday I weigh a parcel. Your friend also (weighs) a parcel. Yesterday you both (weighed) a parcel.
Everyday I lie a little bit. Your friend also (lies) a little bit. Yesterday you both (lied) a little bit.

Experimental items – word-final cluster

Everyday I lick a lollipop. Your friend also (licks) a lollipop. Yesterday you both (licked) a lollipop.
Everyday I cough a lot. Your friend also (coughs) a lot. Yesterday you both (coughed) a lot.
Everyday I wrap a present. Your friend also (wraps) a present. Yesterday you both (wrapped) a present.
Everyday I pack a suitcase. Your friend also (packs) a suitcase. Yesterday you both (packed) a suitcase.
Everyday I hum a tune. Your friend also (hums) a tune. Yesterday you both (hummed) a tune.

References

Au-Yeung, J. & Howell, P. (1998). Lexical and syntactic content and stuttering. *Clinical Linguistic and Phonetics,* **12**, 67-78.

Baayen, H., Piepenbrock, R. & van Rijn, H. (1993). *The CELEX lexical database, CD-ROM*. University of Pennsylvania: Linguistic Data Consortium.

Berko, J. (1958). The child's learning of English morphology. *Word*, **14**, 150-177.

Bernhardt, B. & Stemberger, J. (1998). *Handbook of phonological development*. San Diego: Academic Press.

Bishop, D., North, T. & Donlan, C. (1996). Nonword repetition as a behavioural marker for inherited language impairment: Evidence from a twin study. *Journal of Child Psychiatry*, **4**, 391-403.

Brown, M. C. (2001). *Perl: The complete reference second edition*. New York: McGraw-Hill.

Chiat, S. (1989). The relation between prosodic structure, syllabification and segmental realization: evidence from a child with fricative stopping. *Clinical Linguistics and Phonetics*, **3**, 223-242.

Conture, E.G. (1990). *Stuttering*. Englewood Cliffs, NJ: Prentice-Hall.

Demuth, K. (2004). Production approaches to stuttering. *Stammering Research*, **1**, 297-298.

Dworzynski, K. & Howell, P. (2004). Predicting stuttering from phonetic complexity in German. *Journal of Fluency Disorders*, **29**, 149-173.

Gallon, N., J. Harris & van der Lely, H.K.J. (submitted). Nonword repetition: An investigation of phonological complexity in children with G-SLI.

Gathercole, S. &, Baddeley, A. (1990). Phonological memory deficits in language disordered children: Is there a connection? *Journal of Memory and Language*, **29**, 336-360.

Hakim, H.B. & Ratner, N.B. (2004). Nonword repetition abilities in children who stutter: An exploratory study. *Journal of Fluency Disorders,* **29**, 179-199.

Howell, P. Au-Yeung, J. (1995). The association between stuttering, Brown's factors and phonological categories in child stutterers ranging in age between 2 and 12 years. *Journal of Fluency Disorders*, **20**: 331-344.

Howell, P., Au-Yeung, J. & Sackin, S. (1999). Exchange of stuttering from function words to content words with age. *Journal of Speech, Language and Hearing Research*, **42**, 345-354.

Howell, P., Au-Yeung, J. & Sackin, S. (2000). Internal structure of content words leading to lifespan differences in stuttering. *Journal of Fluency Disorders*, **25**, 1-20.

Howell, P., Au-Yeung, J., Yaruss, S. & Eldridge, K. (submitted). Phonetic difficulty and stuttering in English. *Journal of Fluency Disorders*.

Howell, P., & Dworyznski, K. (in press). Planning and execution processes in speech control by fluent speakers and speakers who stutter. *Journal of Fluency Disorders*.

Howell, P. & Huckvale, M. (2004). Facilities to assist people to research into stammered speech. *Stammering Research*, **1**, 130-242.

Hubbard C.P. & Prins D. (1994). Word familiarity, syllabic stress pattern, and stuttering. *Journal of Speech and Hearing Research*, **37**, 564-571.

Jakielski, K.J. (1998). *Motor organization in the acquisition of consonant clusters*. Dissertation/PhD thesis, University of Texas Austin.

Kadi-Hanifi, K., & Howell, P. (1992). Syntactic analysis of the spontaneous speech of normally fluent and stuttering children. *Journal of Fluency Disorders*, **17**, 151-170.

Labov, W. (1978). *Sociolinguistic patterns*. Oxford: Blackwell.

Leonard, L.B., Eyer, J.A., Bedore, L.M. & Grela, B.G. (1997). Three accounts of the grammatical morpheme difficulties of English-speaking children with Specific Language Impairment. *Journal of Speech, Language and Hearing Research*, **40**, 741-753.

Marshall, C.R. (2004). *The morpho-phonlogical interface in children with Specific Language Impairment*. Unpublished doctoral dissertation, University of London.

Marshall, C.R., Harris, J. & van der Lely, H.K.J. (2003). The nature of phonological representations in children with Grammatical-Specific Language Impairment (G-SLI). In D. Hall, T. Markopoulos, A. Salamoura & S. Skoufaki (eds) *Proceedings of the University of Cambridge First Postgraduate Conference in Language Research,* **1,** 511-517

Marton, K. & Schwartz, R. (2003). Working memory capacity and language processes in children with Specific Language Impairment. *Journal of Speech, Language and Hearing Research*, **46**, 1138-1153.

Natke, U., Sandreiser, P., van Ark, M., Pietrowski, R. & Kalveram, K.T. (2004). Linguistic stress, within-word position and grammatical class in relation to early childhood stuttering. *Journal of Fluency Disorders,* **29**, 109-122.

Packman, A., Onslow, M., Richard, F. & van Doorn, J. (1996). Syllabic stress and variability: A model of stuttering. *Clinical Linguistics and Phonetics*, **10**, 235-263.

Rice, M.L. & Wexler, K. (2001). *Test of Early Grammatical Impairment*. Psychological Corporation.

Rice, M.L., Wexler, K. & Cleave, P.L. (1995). Specific Language Impairment as a period of extended optional infinitive. *Journal of Speech and Hearing Research*, **38**, 850-863.

Roy, P. & Chiat, S. (2004). A prosodically controlled word and nonword repetition task for 2- to 4-year-olds: evidence from typically developing children. *Journal of Speech, Language, and Hearing Research,* **47**, 223-234.

Savage, C. & Lieven, E. (2004). Can the Usage-based approach to language development be applied to analysis of developmental stuttering? *Stammering Research*, **1**, 83-111.

Silverman, S., & Bernstein Ratner, N. (1997). Syntactic complexity, fluency, and accuracy of sentence imitation in adolescents. *Journal of Speech, Language, and Hearing Research*, **40**, 95-106.

Throneburg, R.N., Yairi, E. & Paden, E.P. (1994). Relation between phonologic difficulty and the occurrence of disfluencies in the early stage of stuttering. *Journal of Speech and Hearing Research*, **37**, 504-509.

van der Lely, H.K.J. & Harris, J. (1999). *The Test of Phonological Structure*. Unpublished test available from the first author, Centre for Developmental Language Disorders and Cognitive Neuroscience, Department of Human Communication Science, University College London, London UK.

van der Lely, H.K.J. & Ullman, M. (2001). Past tense morphology in specifically language impaired and normally developing children. *Language and Cognitive Processes*, **16**: 177-217

Weiss, A. & Jakielski, K. (2001). Phonetic complexity measurement and prediction of children's disfluencies: a preliminary study. *Proceedings of the 4th International Speech Motor Control Conference*, Nijmegen, the Netherlands: Univeristy of Nijmegen, Nijmegen Medical Centre.

Wingate, M. (1982). Early position and stuttering occurrence. *Journal of Fluency Disorders*, **7**, 243-258.

Wingate, M. (1988). *The structure of stuttering: A psycholinguistic analysis*. Berlin: Springer-Verlag.

Yairi, E., & Ambrose, N. G. (2004). *Early childhood stuttering*. AustinTX: Pro-Ed.

Stammering Research
Index to Volume 1, 2004-2005

Author index

Everard, R. (2004). Commentary on Partnerships between Clinicians, Researchers and People who Stutter in the Evaluation of Stuttering Treatment Outcomes. *Stammering Research,* **1,** 24-25.

Furnham, A., & Davis, S. (2004). Involvement of social factors in stuttering: A review and assessment of current methodology. *Stammering Research,* **1,**112-122

Gershkoff-Stowe, L. (2004). Commentary on ‘Can the Usage-Based Approach to Language Development be Applied to Analysis of Developmental Stuttering?’ by C.Savage and E.Lieven. *Stammering Research,* **1,** 101-102.

Ghazi, S., & Lickley, R. (2004). Bilingual issues in stammering intervention: an exploratory study. *Stammering Research,* **1,** 327

Gooding, S., & Davis, S. (2004). Indications by parents of ‘environmental’ factors that differ between children who stutter and controls. *Stammering Research,* **1,** 21-23

Howell, P. (2004). Effects of delayed auditory feedback and frequency-shifted feedback on speech control and some potentials for future development of prosthetic aids for Stammering. *Stammering Research,* **1***,* 31-46.

Howell, P. (2004). Response to commentaries on ‘Effects of delayed auditory feedback and frequency-shifted feedback on speech control and some potentials for future development of prosthetic aids for stammering’. *Stammering Research,* **1,** 68-77.

Howell, P. (2005). The effect of using time intervals of different length on judgements about stuttering. *Stammering Research,* **1,** 364-374.

Howell, P., Davis, S., Bartrip, J. & Wormald, L. (2004).Effectiveness of frequency shifted feedback at reducing disfluency for linguistically easy, and difficult, sections of speech (original audio recordings included). *Stammering Research,* **1,** 309-315.

Howell, P., & Huckvale, M. (2004). Facilities to assist people to research into stammered speech. *Stammering Research,* **1,** 130-242

Hussain, K., Khan, S., Howell, P., & Davis, S. (2004). Do children who stutter have word finding difficulties? *Stammering Research,* **1,** 326

Joukov, S. (2004). Trial software for frequency shifted and delayed auditory feedback *Stammering Research,* **1,** 316-325.

Levine, S. Z., Petrides, K. V., Davis, S., Jackson, C. J., & Howell, P. (2005). The use of structural equation modeling in stuttering research: Concepts and directions. *Stammering Research,* **1,** 344-363

Lowitt, A., & Brendel, B. (2004). The response of patients with Parkinson's Disease to DAF and FSF. *Stammering Research,* **1,** 58-61

Manning, W. H. (2004). The social phenomena of stuttering. *Stammering Research,* **1,** 123-124.

Marshall, C. (2005). The impact of word-end phonology and morphology on stuttering. *Stammering Research,* **1,** 375-391.

Miller, D. (2004). Effects of using an Edinburgh masker for a period of 25 years. *Stammering Research,* **1,** 67.

Neumann, K., & Euler, H. (2004). Are alerts sufficient to smooth speech? *Stammering Research,* **1,** 48-50.

Osborne, V. (2004). Comments on the implications of Yaruss and Quesal (2004) on seeking, and measuring outcomes of therapy from the perspective of an adult covert stammerer. *Stammering Research,* **1,** 26-27.

Pfordresher, P. (2004). Can altered auditory information affect planning? Evidence from music performance. *Stammering Research,* **1,** 51-53

Reed, P. (2004). Altered Recurrent Auditory Information and Stammering: Contributions from Learning Theory. *Stammering Research,* **1,** 55-57

Ryan, B. (2004). Programmed Stuttering Therapy for Children and Adults using ARAI. *Stammering Research,* **1,** 62

Sage, R. (2004). How interviews with adults who stammer inform research directions concerning what we think about our stammer: Participants views from the Apple House Fluency Course, Oxford. *Stammering Research,* **1,** 271-289.

Savage, C. (2004). Responses to commentators. *Stammering Research,* **1,** 107-111.

Savage, C., & Lieven, E. (2004). Can the Usage-Based approach to language development be applied to analysis of developmental stuttering? *Stammering Research,* **1,** 83-100

Savage, C., & Lieven, E. (2004). Can the Usage-Based approach to language development be applied to analysis of developmental stuttering? *Stammering Research,* **1,** 299-300

Schwartz, M. (2004). Commentary on 'Can the Usage-Based approach to language development be applied to analysis of developmental stuttering?' by C.Savage and E.Lieven. *Stammering Research,* **1,** 106.

Skljarov, O. P. (2004). Commentary on 'Can the Usage-Based approach to language development be applied to analysis of developmental stuttering?' by C.Savage and E.Lieven. *Stammering Research,* **1,** 103-105.

Skljarov, O. P. (2004). Commentary on 'Involvement of social factors in stuttering: A review and assessment of current methodology' by A. Furnham and S. Davis, *Stammering Research,* **1,** 127.

Skotko, J. (2004). Experience of a speech pathologist providing clients with the SpeechEasy^{TM} device. *Stammering Research,* **1,** 63-65.

Stenner, P. (2004). On reaching the parts that quantitative researchers cannot: Commentary on Sage, R. (2004) "How interviews with adults who stammer inform research directions". *Stammering Research,* **1,** 290-292

Vanryckeghem, M. (2004). Commentary on "Involvement of social factors in stuttering: A review and assessment of current methodology" By A. Furnham and S. Davis. *Stammering Research,* **1,** 304-306

Wade, J. (2004). Extending the scope of stuttering research and treatment. *Stammering Research,* **1,** 125-126.

Wormald, L., Davis, S., & Howell, P. (2004). The Family History of Stuttering in Recovered and Persistent Groups. *Stammering Research,* **1,** 326

Yaruss, J. S.,& Quesal, R. W. (2004). Partnerships between Clinicians, Researchers, and People Who Stutter in the Evaluation of Stuttering Treatment Outcomes. *Stammering Research,* **1**, 1-15.

Yaruss, J. S., & Quesal, R. W. (2004). The Best Way For Clinicians To Learn What Their Clients Want From Treatment Is To Ask Them What They Want From Treatment. *Stammering Research,* **1,** 28-30.

Subject index

Target articles

Commentaries and response to commentaries

Yaruss and Quesal

Gooding, S., & Davis, S. (2004). Indications by parents of 'environmental' factors that differ between children who stutter and controls. *Stammering Research,* **1,** 21-23

Everard, R. (2004). Commentary on Partnerships between Clinicians, Researchers and People who Stutter in the Evaluation of Stuttering Treatment Outcomes. *Stammering Research,* **1,** 24-25.

Osborne, V. (2004). Comments on the implications of Yaruss and Quesal (2004) on seeking, and measuring outcomes of therapy from the perspective of an adult covert stammerer. *Stammering Research,* **1,** 26-27.

Yaruss, J. S., & Quesal, R. W. (2004). The Best Way For Clinicians To Learn What Their Clients Want From Treatment Is To Ask Them What They Want From Treatment. *Stammering Research,* **1,** 28-30.

Howell

Neumann, K., & Euler, H. (2004). Are alerts sufficient to smooth speech? *Stammering Research,* **1,** 48-50.

Pfordresher, P. (2004). Can altered auditory information affect planning? Evidence from music performance. *Stammering Research,* **1,** 51-53

Davis, S. (2004). What sort of cerebellar processing problem do people who stutter have? *Stammering Research,* **1,** 54

Reed, P. (2004). Altered Recurrent Auditory Information and Stammering: Contributions from Learning Theory. *Stammering Research,* **1,** 55-57

Lowitt, A., & Brendel, B. (2004). The response of patients with Parkinson's Disease to DAF and FSF. *Stammering Research,* **1,** 58-61

Ryan, B. (2004). Programmed Stuttering Therapy for Children and Adults using ARAI. *Stammering Research,* **1,** 62

Skotko, J. (2004). Experience of a speech pathologist providing clients with the SpeechEasy™ device. *Stammering Research,* **1,** 63-65.

Bartles, S., & Ramig, P. R. (2004). Clinical Research into use of the SpeechEasy™ device. *Stammering Research,* **1,** 66.

Miller, D. (2004). Effects of using an Edinburgh masker for a period of 25 years. *Stammering Research,* **1,** 67.

Howell, P. (2004). Response to commentaries on ‘Effects of delayed auditory feedback and frequency-shifted feedback on speech control and some potentials for future development of prosthetic aids for stammering’. *Stammering Research,* **1,** 68-77.

Savage and Lieven

Gershkoff-Stowe, L. (2004). Commentary on ‘Can the Usage-Based Approach to Language Development be Applied to Analysis of Developmental Stuttering?’ by C.Savage and E.Lieven. *Stammering Research,* **1,** 101-102.

Skljarov, O. P. (2004). Commentary on ‘Can the Usage-Based Approach to Language Development be Applied to Analysis of Developmental Stuttering?’ by C.Savage and E.Lieven. *Stammering Research,* **1,** 103-105.

Schwartz, M. (2004). Commentary on ‘Can the Usage-Based Approach to Language Development be Applied to Analysis of Developmental Stuttering?’ by C.Savage and E.Lieven. *Stammering Research,* **1,** 106.

Anderson, J. D., & Musolino, J. (2004). How Useful is the Usage Based Approach to Stuttering Research? *Stammering Research,* **1,** 295-296

Demuth, K. (2004). Production approaches to stuttering. *Stammering Research,* **1,** 297-298

Savage, C. (2004). Responses to commentators. *Stammering Research,* **1,** 107-111.

Savage, C., & Lieven, E. (2004). Can the Usage-Based Approach to Language Development be Applied to Analysis of Developmental Stuttering? *Stammering Research,* **1,** 299-300

Furnham and Davis

Manning, W. H. (2004). The social phenomena of stuttering. *Stammering Research,* **1,** 123-124.

Wade, J. (2004). Extending the scope of stuttering research and treatment. *Stammering Research,* **1,** 125-126.

Skljarov, O. P. (2004). Commentary on ‘Involvement of social factors in stuttering: A review and assessment of current methodology’ by A. Furnham and S. Davis, *Stammering Research,* **1,** 127.

Blood, G. & Blood, I. M. (2004). Social Factors in Stuttering: A Call for Rigorous, Accountable and Timely Research. *Stammering Research,* **1,** 301-303

Vanryckeghem, M. (2004). Commentary on "Involvement of social factors in stuttering: A review and assessment of current methodology" By A. Furnham and S. Davis. *Stammering Research,* **1,** 304-306

Davis, S., & Furnham, A. (2004). Authors response to commentaries on 'Involvement of social factors in stuttering: A review and assessment of current methodology'. *Stammering Research,* **1,** 128-129.

Davis, S., & Furnham, A. (2004). Involvement of social factors in stuttering: A review and assessment of current methodology. *Stammering Research,* **1,** 307-308

Sage

Stenner, P. (2004). On reaching the parts that quantitative researchers cannot: Commentary on Sage, R. (2004) "How interviews with adults who stammer inform research directions". *Stammering Research,* **1,** 290-292

Research data, software and analysis section

Howell, P. (2005). The effect of using time intervals of different length on judgements about stuttering. *Stammering Research,* **1,** 364-374.

Howell, P., & Huckvale, M. (2004). Facilities to assist people to research into stammered speech. *Stammering Research,* **1,** 130-242

Howell, P., Davis, S., Bartrip, J. & Wormald, L. (2004). Effectiveness of frequency shifted feedback at reducing disfluency for linguistically easy, and difficult, sections of speech (original audio recordings included). *Stammering Research,* **1,** 309-315.

Joukov, S. (2004). Trial software for frequency shifted and delayed auditory feedback *Stammering Research,* **1,** 316-325.

Marshall, C. (2005). The impact of word-end phonology and morphology on stuttering. *Stammering Research,* **1,** 375-391.

Abstracts of reports from BSA vacation studentships 2004

Wormald, L., Davis, S., & Howell, P. (2004). The Family History of Stuttering in Recovered and Persistent Groups. *Stammering Research,* **1,** 326

Hussain, K., Khan, S., Howell, P., & Davis, S. (2004). Do children who stutter have word finding difficulties? *Stammering Research,* **1,** 326

Ghazi, S., & Lickley, R. (2004). Bilingual issues in stammering intervention: an exploratory study. *Stammering Research,* **1,** 327

Title index

Commentary on 'Involvement of social factors in stuttering: A review and assessment of current methodology' by A. Furnham and S. Davis. Skljarov, O. P. (2004). *Stammering Research,* **1,** 127.

Commentary on "Involvement of social factors in stuttering: A review and assessment of current methodology" By A. Furnham and S. Davis. Vanryckeghem, M. (2004). *Stammering Research,* **1,** 304-306

Commentary on Partnerships between Clinicians, Researchers, and People Who Stutter in the Evaluation of Stuttering Treatment Outcomes by J. Scott Yaruss and Robert W. Quesal. Dahm, B. (2004). *Stammering Research,* **1,** 16-17.

Commentary on Partnerships between Clinicians, Researchers and People who Stutter in the Evaluation of Stuttering Treatment Outcomes. Everard, R. (2004). *Stammering Research,* **1,** 24-25.

Comments on the implications of Yaruss and Quesal (2004) on seeking, and measuring outcomes of therapy from the perspective of an adult covert stammerer. Osborne, V. (2004). *Stammering Research,* **1,** 26-27.

Do children who stutter have word finding difficulties? Hussain, K., Khan, S., Howell, P., & Davis, S. (2004). *Stammering Research,* **1,** 326

Effectiveness of frequency shifted feedback at reducing disfluency for linguistically easy, and difficult, sections of speech (original audio recordings included). Howell, P., Davis, S., Bartrip, J. & Wormald, L. (2004).*Stammering Research,* **1,** 309-315.

Effects of delayed auditory feedback and frequency-shifted feedback on speech control and some potentials for future development of prosthetic aids for Stammering. Howell, P. (2004). *Stammering Research,* **1**, 31-46.

Effects of using an Edinburgh masker for a period of 25 years. Miller, D. (2004). *Stammering Research,* **1,** 67.

Elements of statistical treatment of speech and hearing science data. Davis, A., & Howell, P. (2005). *Stammering Research,* **1,** 333-343

Experience of a speech pathologist providing clients with the SpeechEasy™ device. Skotko, J. (2004). *Stammering Research,* **1,** 63-65.

Extending the scope of stuttering research and treatment. Wade, J. (2004). *Stammering Research,* **1,** 125-126.

Facilities to assist people to research into stammered speech. Howell, P., & Huckvale, M. (2004). *Stammering Research,* **1,** 130-242